

---

# MEGI

**MESTRADO**

Estatística e Gestão de Informação

---

---

*Análise do Mercado Financeiro baseada em Análise  
Técnica com Self-Organizing Maps*

---

Gisele Cristina Panosso

---

Dissertação apresentada como requisito parcial para  
obtenção do grau de Mestre em Estatística e Gestão de  
Informação

Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

## **ANÁLISE DO MERCADO FINANCEIRO BASEADA EM ANÁLISE TÉCNICA COM SELF-ORGANIZING MAPS**

Por

Gisele Cristina Panosso

Dissertação apresentada como requisito parcial para a obtenção do grau de Mestre em  
Estatística e Gestão de Informação, Especialização em Business Intelligence

Orientador/Coorientador: Prof. Fernando Bação  
Coorientador: Prof. Nuno Cavalheiro Marques

## **RESUMO**

O investimento no mercado de ações é um dos tipos mais populares do mercado financeiro. O objetivo dos investidores e analistas financeiros é maximizar os lucros e minimizar as perdas, para tanto uma das técnicas mais utilizadas é a Análise Técnica. Tradicionalmente análises de ações são baseadas no histórico de preços com objetivo de prever o futuro. A abordagem desta pesquisa foi empregar indicadores da Análise Técnica com Mapas Auto-Organizáveis com o objetivo de fazer segmentação e posterior análise das variáveis mais significativas na descrição do comportamento do movimento dos preços. O índice utilizado foi *Dow Jones Industrial Average* no período de 2001 a 2011.

## **PALAVRAS-CHAVE**

Mapas Auto-Organizáveis; Análise Técnica; Índice de Hurst; Mercado Financeiro

## **ABSTRACT**

Investments in the stock market are one of the most popular types of financial market. The objective of financial analysts and investors is maximizing profits and minimize the losses, to that end one of the most used techniques is Technical Analysis. Traditionally analyses are based on historical stock prices in order to predict the future. The approach of this research was to employ indicators of Technical Analysis with Self-Organizing Maps in order to do segmentation and subsequent analysis of the most significant variables in description of the behavior of the price movement. The index used was Dow Jones Industrial Average from 2001 to 2011.

## **KEYWORDS**

Self-Organizing Maps, Technical Analysis, Hurst Index; Stock Market

## ÍNDICE

1. Introdução .....	9
1.1. Objetivos do estudo .....	10
1.2. Importância e relevância do estudo .....	10
2. Revisão da Literatura .....	12
2.1. Teorias económicas.....	12
2.2. Os Mapas Auto-Organizáveis .....	16
2.3. Trabalhos relacionados .....	18
3. Metodologia .....	22
3.1. Classificação da Pesquisa .....	22
3.2. Fonte de Dados .....	22
3.3. Variáveis .....	23
3.4. Redução das Variáveis .....	29
3.5. Análise de Subconjuntos .....	30
3.6. Processamento e Análise de Dados .....	31
4. Resultados e Discussão .....	37
4.1. Redução das variáveis .....	37
4.2. Tamanho do Mapa .....	41
4.3. Topologia e Normalização.....	42
4.4. Resultados Self-Organizing Map .....	43
4.4.1. Resultados Parciais .....	43
4.4.2. Período de 2001 a 2011 .....	48
4.4.3. Período de 2002 a 2007 .....	56
4.5. Validação dos Resultados.....	61
5. Conclusões, Limitações e Recomendações.....	64
6. Bibliografia .....	66
7. Anexos .....	70
7.1. Cálculo do Índice de Hurst .....	70
7.2. Package Oracle utilizado para cálculo das variáveis .....	71

## ÍNDICE DE FIGURAS

Figura 2.1 – Mapa Auto-Organizável .....	17
Figura 3.1 - Períodos de estudo .....	30
Figura 3.2 - Topologia de vizinhança .....	32
Figura 3.3 - Formas da matriz .....	32
Figura 3.4 - Funções de vizinhança .....	33
Figura 3.5 – U-Matrix <i>World Poverty Map</i> .....	35
Figura 3.6 – Segmentação GeoSOM Suite .....	36
Figura 4.1 - Índice de Hurst Médio e Longo Prazo: Período 2001 a 2011 .....	37
Figura 4.2 - Árvore clusters de variáveis .....	38
Figura 4.3 - Gráfico de Erros x Número de neurónios – 2001 a 2011 .....	41
Figura 4.4 - Gráfico de Erros x Número de neurónios – 2002 a 2007 .....	42
Figura 4.5 - Gráfico de Erros x Número de neurónios – 2002 a 2011 .....	44
Figura 4.6 - Umat e Umat com <i>Hits</i> com Var. Preço: Período 2002 a 2011 .....	45
Figura 4.7 – Matriz de <i>Labels</i> com Var. Preço: Período 2002 a 2011 .....	46
Figura 4.8 - Umat e Umat com <i>Hits</i> sem Var. Preço: Período 2002 a 2011.....	46
Figura 4.9 – Matriz de <i>Labels</i> sem Var. Preço: Período 2002 a 2011 .....	47
Figura 4.10 - Umat e Umat com Hits: Período 2001 a 2011 .....	48
Figura 4.11 – Matriz de Labels: Período 2001 a 2011 .....	49
Figura 4.12 - Component Planes: Período 2001 a 2011 .....	51
Figura 4.13 - Índice de Hurst Médio Prazo: Período 2001 a 2011 .....	52
Figura 4.14 - Análise de Cluster Vales: Período 2001 a 2011 .....	54
Figura 4.15 - Análise de Cluster Topos: Período 2001 a 2011 .....	55
Figura 4.16 - UMAT e UMAT com Hits: Período 2002 a 2007 .....	56
Figura 4.17 – Matriz de Labels: Período 2002 a 2007 .....	57
Figura 4.18 - Component Planes: Período 2002 a 2007 .....	57
Figura 4.19 - Volume Médio: Tendências de baixa, alta e inversão do movimento .....	58
Figura 4.20 - Segmentação Mapa 2002 a 2007 .....	59
Figura 4.21 - Segmentação Mapa 2002 a 2007 .....	60
Figura 4.22 – Evolução do capital líquido no fechamento mensal.....	62

## ÍNDICE DE TABELAS

Tabela 3.1 - Tabela das Variáveis.....	28
Tabela 3.2 - Fórmula tamanho do mapa .....	31
Tabela 4.1 - Tabela Estratégias de Investimento.....	61

## LISTA DE SIGLAS E ABREVIATURAS

<b>BMU</b>	Best Matching Unit
<b>DJI</b>	Dow Jones Industrial Average
<b>ESOM</b>	Emergent Self-Organizing Maps
<b>HEM</b>	Hipótese de Eficiência dos Mercados
<b>SOM</b>	Self-Organizing Map
<b>UMAT</b>	Unified Distance Matrix



## 1. INTRODUÇÃO

O mercado de ações representa uma excelente alternativa para os investidores, uma vez que os retornos dos investimentos geralmente apresentam altas taxas de rentabilidade, tratando-se no entanto de um tipo de investimento de alto risco onde não há garantias de que ao final de um determinado período o lucro seja obtido.

Neste mercado a previsão dos preços bem como encontrar o momento certo para investir ou sair do mercado são certamente de grande interesse para os investidores. Para atingir estes objetivos as principais abordagens empregadas são a Análise Técnica e a Fundamental.

A Análise Técnica tem como premissa o preço das ações e as tendências de movimento dos preços, sendo a principal característica tentar antecipar a direção do próximo movimento de preços de uma ação, a fim de possibilitar o investimento de curto ou longo prazo em conformidade, sinalizar o momento para tomar decisões de compra, venda ou manter posições, bem como melhorar a estimativa de tempo dos movimentos de preço das ações baseando-se em técnicas como gráficos, fórmulas de preços/ volume e observação de faixas de preços das negociações (Thomsett, 2007).

A premissa subjacente da Análise Técnica é encontrada na Teoria Dow. Segundo a teoria os preços das ações tendem a agir em concerto, isto é, quando a tendência é de subida há tendências do mercado global nessa direção e vice-versa. Assim, um número limitado de líderes de mercado pode ser identificado e demais ativos financeiros seguirão a liderança estabelecida por essas empresas influentes (Rhea, 1932). Além disso, a Análise Técnica sustenta-se na hipótese de que os padrões passados de comportamento de preços em cada ativo tenderão a se repetir no futuro.

Já a Análise Fundamental é uma análise baseada em fatores económicos, projeções de comportamento futuro, condições de oferta e procura de bens e serviços e nos fundamentos da economia, a qual pretende definir a saúde financeira de empresas no interior de suas áreas de atuação. Ao contrário da Análise Técnica, a Análise Fundamental não se baseia no estudo das cotações de bolsa, ela interpreta dados fundamentais obtidos no balanço de uma empresa (Deschatre G. , 2009).

A partir das informações fundamentais resultantes da análise de balanços, análise setorial, económica, estudo da gestão da empresa e tendências de consumo são traçadas projeções de resultados para determinar preços justos para as ações das empresas (Cerbasi, 2008). Com base nesse valor intrínseco (resultado da Análise Fundamental) e a comparação do valor de mercado das ações é que se pode determinar se um ativo está supervalorizado ou subvalorizado.

Há ainda outra teoria que contrapõe as ideias defendidas pela Análise Técnica, é conhecida como a Teoria dos Passeios Aleatórios. Segundo esta teoria as sucessivas mudanças de preço são independentes, ou seja, o passado não pode ser usado para prever o futuro (Fama, 1965).

Contudo, verifica-se que há controvérsias entre as teorias existentes para análise e previsão dos movimentos no mercado de ações, o que torna ainda mais difícil o trabalho dos analistas financeiros e investidores.

Esta pesquisa fundamenta-se nas teorias da Análise Técnica pois se utilizará de seus indicadores para criar um modelo de segmentação com mapas auto-organizáveis (*Self-Organizing Maps* – SOM) e investigar a importância destes indicadores na discriminação dos movimentos dos preços.

No capítulo 2 Revisão da Literatura as teorias do mercado financeiro serão discutidas com maiores detalhes.

### **1.1. OBJETIVOS DO ESTUDO**

O objetivo geral deste trabalho será experimentar a combinação de indicadores da Análise Técnica para identificar padrões utilizando mapas auto-organizáveis (SOM's). Baseado no comportamento histórico do mercado, identificar quais variáveis são mais significativas para descrever o comportamento dos ativos financeiros e capazes de sinalizar momentos de estabilidade ou de crise financeira.

Os objetivos secundários são:

- Identificar quais os intervalos são mais expressivos para cálculo das variáveis, ou seja, na Análise Técnica muitas variáveis são calculadas por intervalos de tempo: curto, médio, longo prazo. É importante determinar intervalos que tornem as variáveis mais expressivas.
- Avaliar as correlações entre as variáveis, com o objetivo de eliminar variáveis com alta correlação e não prejudicar a eficiência do modelo.
- Validar os resultados obtidos ao simular investimentos em momentos mais oportunos de acordo com os mapas auto-organizáveis.

### **1.2. IMPORTÂNCIA E RELEVÂNCIA DO ESTUDO**

O objetivo dos investidores e analistas financeiros é maximizar os lucros. Para tanto é essencial utilizar todas as técnicas possíveis para prever o comportamento dos mercados financeiros.

Entretanto, em tempos onde as crises financeiras tornam-se cada vez mais constantes e provocam grandes quedas no mercado de investimentos, as decisões de aplicação em títulos são cada vez mais difíceis e arriscadas, seja para novos investidores ou analistas financeiros experientes.

Dessa forma, identificar indicadores que possam sinalizar momentos de estabilidade e de crise financeira será de extrema importância e até possibilitará a construção de ferramentas de apoio à decisão que classifiquem o momento atual e sinalizem o risco de um investimento.

Consequentemente, uma ferramenta de apoio à decisão confiável poderá evitar grandes perdas aos investidores e falência de empresas de investimento, pois permitirá saber até que ponto se pode confiar nas tendências sugeridas pelos indicadores da Análise Técnica e/ou outras ferramentas de previsão.

## 2. REVISÃO DA LITERATURA

Esta secção foi dividida em 3 partes. A primeira designada Teorias Económicas apresenta as principais teorias relacionadas com o comportamento dos ativos no mercado financeiro. Na segunda secção dedicada Os Mapas Auto-Organizáveis, é feito um breve resumo do algoritmo que será utilizado para processamento dos dados e por fim, na terceira secção são apresentados alguns dos principais trabalhos relacionados com o tema deste trabalho.

### 2.1. TEORIAS ECONÓMICAS

Charles Henry Dow foi um jornalista americano muito importante para o mercado financeiro. Em 1889 começou a publicação de um pequeno jornal chamado *The Wall Street Journal*. Entre 1899 e 1902, Dow escreveu para seu jornal uma série de editoriais sobre o mercado financeiro, editoriais que forneceram a base para os princípios subjacentes da Teoria de Dow.

Em 1896 Dow introduziu pela primeira vez sua média de stock industrial como um índice de 12 empresas, o *Dow Jones Industrial Average* (DJI), índice que até hoje é um dos principais índices da bolsa de valores dos Estados Unidos da América.

A Teoria de Dow foi publicada após a sua morte por William P. Hamilton, sucessor de Dow no *The Wall Street Journal*. Hamilton publicou muitas das observações e filosofia de Dow num livro chamado "*The Stock Market Barometer*" (Dow, Russell, Carlson, Shread & Sether, 2009).

A Teoria de Dow pode ser considerada como filosofia do mercado e foi a base da Análise Técnica. A Teoria de Dow apresenta os seguintes pressupostos:

As médias: as médias representam tudo o que se sabe e tudo o que pode ser previsto por instituições sobre questões financeiras. Com efeito, as médias refletem com precisão cada fonte de informação importante e que tem algum significado no mercado.

As médias são compostas por três movimentos de preços: a tendência primária, reação secundária e flutuações diárias. Tendências primárias são conhecidas como o mercado do Touro e do Urso, ou seja, o Touro representa os investidores que querem elevar o preço dos ativos e o Urso representa os investidores que têm interesse que o preço caia. Essa tendência pode durar de menos de um ano até vários anos. Reações secundárias representam as correções da tendência primária. Já as flutuações diárias

refletem as pequenas variações da tendência secundária e, de acordo com a Teoria de Dow, oferecem pouco no poder de previsão. (Dow, Russell, Carlson, Shread & Sether, 2009).

A Análise Técnica, com as premissas da Teoria de Dow, utiliza-se de gráficos para analisar o histórico do movimento dos mercados e alguns indicadores matemáticos para tentar prever as tendências do comportamento futuro. O principal pressuposto da Análise Técnica é a de que os padrões passados de preço tendem a se repetir no futuro.

Análise Técnica e Fundamental não são apenas duas abordagens da mesma questão, são totalmente diferentes em sua natureza. Análise Técnica foca nas tendências de preço das ações enquanto Análise Fundamental trata de resultados financeiros corporativos.

A Análise Fundamental acredita que os resultados financeiros são os únicos meios confiáveis para estabelecer o valor de uma empresa. A tendência de preço, na visão fundamental, é de curto prazo, caótica e incerta, causada por muitos fatores momentâneos conflitantes e imprecisos. A oferta e procura imediata dentro do mercado de ações é ilógica, os movimentos dos preços reagem a notícias e boatos, e muito do movimento diário é randômico ou artificial (Thomsett M. C., 2006).

Esta visão da Análise Fundamental vai de encontro à outra teoria conhecida como Teoria do Passeio Aleatório (*Random Walks*), publicada por Fama em 1965. Segundo esta teoria, o caminho futuro do preço de um título não é mais previsível do que o caminho de uma série acumulada de números aleatórios. Em termos estatísticos a teoria diz que mudanças de preços sucessivos são independentes, como a distribuição de variáveis aleatórias. Isto implica simplesmente que uma série de variações de preços não tem memória, isto é, o passado não pode ser usado para prever o futuro de forma significativa (Fama, 1965).

Ao comparar a Teoria de Dow com a Teoria dos Passeios Aleatórios, percebe-se que elas são opostas em muitos aspectos. A única semelhança é que ambas estão relacionadas com o preço dos títulos. Enquanto a Teoria de Dow acredita que as tendências são previsíveis, a Teoria dos Passeios Aleatórios diz que o movimento dos preços é imprevisível e até arbitrário.

Muito relacionada com a Teoria dos Passeios Aleatórios está a consequente Hipótese de Eficiência dos Mercados (HEM) publicada por Fama (1970). Segundo esta teoria, o preço atual de um título reflete integralmente a informação disponível sobre

uma empresa e os preços mudam instantaneamente para refletir novas informações. As informações são divididas em três subconjuntos de acordo com a sua relevância: fraca, semiforte e forte.

Primeiro, a forma fraca assume que os preços incorporam o conjunto de informações sobre o histórico dos preços. A forma semiforte assume que os preços refletem toda a informação pública disponível. Por fim, a forma forte acredita que os preços incorporam qualquer informação relevante para a formação de preços, mesmo informações secretas ou privilegiadas a que apenas certos investidores ou grupos tenham acesso monopolista.

Juntas, as duas hipóteses (Teoria dos Passeios Aleatórios e Hipótese de Eficiência dos Mercados) constituem o Modelo de Passeio Aleatório (Fama, 1970).

A interpretação da HEM é a de que num mercado eficiente todas as oportunidades de lucros excessivos são anuladas pela concorrência. Porém, a validade da HEM tem sido questionada e até posta em causa devido às recorrentes crises económicas. Basu (1977) na sua validação à HEM concluiu que o comportamento dos preços de títulos estudados num período de 14 anos é, talvez, não completamente descrito pela Hipótese de Eficiência dos Mercados. Contrariamente à crença de que a informação publicamente disponível é instantaneamente incorporada pelos preços dos títulos, parece haver defasagens e atritos no processo de ajuste. Defensores da hipótese contestam que a HEM não consegue garantir a estabilidade de um mercado se a informação publicamente disponível for instável, o mercado também pode acompanhar a instabilidade das informações.

De acordo com as teorias apresentadas, pode-se perceber que há muitas controvérsias em relação ao comportamento dos preços no mercado financeiro. De fato, os mercados são bastantes complexos e não existe nenhuma fórmula infalível para prever os movimentos futuros. No âmbito deste trabalho não cabe defender ou provar nenhuma das teorias apresentadas, mesmo porque não há consenso no próprio meio académico em favor de uma ou contrária à outra. Este trabalho pretende investigar um conjunto de variáveis provenientes da Análise Técnica por esta se apresentar como uma das principais estratégias de investimentos utilizada nos mercados financeiros no último centenário.

Os principais conceitos da Análise Técnica importantes para a compreensão deste trabalho são preço, volume, médias móveis e índice de Hurst. Os analistas técnicos acreditam que dados de preços e volume fornecem indicadores dos futuros

movimentos de preços. Se os mercados são eficientes no sentido de que o preço atual reflete todas as informações, então esta teoria da Análise Técnica é claramente inútil. Mas se o processo pelo qual os preços se ajustam à informação não é imediata, então as estatísticas do mercado podem incorporar informações que ainda não estão refletidas no preço atual (Blume, Easley & O'Hara, 1994).

De fato, a combinação do preço com o volume pode revelar informações valiosas sobre a força dos movimentos. No trabalho desenvolvido por Blume, Easley & O'Hara (1994) eles concluíram que o volume captura informação importante contida nos sinais dos mercados. E uma vez que a existência de rumores pode afetar os preços das ações, o uso da informação revelada pelo volume pode ser particularmente útil para os investidores que operam em mercados voláteis.

A média móvel calculada sobre um índice de ações significa o nível médio do índice sobre um dado intervalo de tempo. Uma média móvel de 52 semanas, por exemplo, rastreia o valor médio do índice sobre as 52 semanas mais recentes. A cada semana, a média móvel é novamente calculada ao retirar a observação mais antiga e adicionar a mais recente. Após um período no qual os preços têm geralmente caído, é suposto que a média móvel fique acima do preço atual pois inclui preços mais antigos e provavelmente mais altos. Segundo as teorias da Análise Técnica, quando o preço de mercado rompe a linha da média móvel por baixo, considera-se um sinal de alta e significa uma mudança de tendência de queda para tendência de alta. Por outro lado, quando os preços estão a subir caem abaixo da média móvel, sendo considerada hora de vender, pois sinaliza nova inversão de tendência para queda dos preços. (Marcus, Bodie & Kane, 2007).

O índice de Hurst envolve uma importante teoria para o desenvolvimento desta pesquisa, que consiste na memória de longo prazo de séries temporais. A primeira estatística foi proposta pelo hidrólogo Inglês Harold Edwin Hurst (1951), que ao estudar projetos de dimensionamento de represas no rio Nilo analisou décadas de registros de cheias e secas, e desenvolveu um método estatístico para resolver o problema da construção de um reservatório ideal. Ele observou correlações nas séries de dados analisadas e seus estudos deram origem à estatística chamada de "Rescaled Range" ou "R/S" estatística. Posteriormente Mandelbrot & Wallis (1969) realizaram estudos com a estatística R/S e classificaram-na como uma estatística muito robusta para testar a presença de memória em séries não-cíclicas. Mandelbrot (1971) também considerou a existência de memória em ativos financeiros, sendo esta a abordagem

que se aplica a este trabalho. Mais detalhes sobre o índice de Hurst são apresentados nos Anexos, capítulo 7.1 Cálculo do Índice de Hurst.

No próximo capítulo (3. Metodologia) serão apresentadas as variáveis e métodos utilizados no desenvolvimento desta pesquisa.

## **2.2. OS MAPAS AUTO-ORGANIZÁVEIS**

As variáveis selecionadas serão estudadas através do uso dos mapas auto-organizáveis, que é um método computacional para a visualização e análise de dados de alta dimensionalidade (Kohonen, 1982).

De acordo com Deboeck & Kohonen (1998), o SOM pertence à classe de métodos de redes neuronais, que são técnicas de regressão não-linear que podem ser treinados para aprender ou encontrar relações entre entradas e saídas ou para organizar os dados, de modo a descobrir padrões ou estruturas desconhecidas, e por este motivo os mapas auto-organizáveis são particularmente interessantes para este trabalho.

Os mapas auto-organizáveis são redes neuronais com alimentação para a frente, ou seja, redes onde a informação se move apenas numa direção, para a frente, a partir das unidades de entrada para as unidades de saída, não existindo ciclos na rede. Utilizam um algoritmo de treino não supervisionado e através de um processo de auto-organização configuram as unidades de saída numa representação topológica dos dados originais. O SOM é capaz de reduzir dados multidimensionais em mapas de baixa dimensão, geralmente bidimensionais.

O SOM é baseado na aprendizagem competitiva não supervisionada. Ele fornece um mapeamento com preservação da topologia do espaço multidimensional, isto significa que um grupo semelhante no espaço de entrada é mapeado para unidades próximas no mapa do SOM (Deboeck & Kohonen, 1998).

Um mapa auto-organizável é composto por duas camadas de unidades de processamento: uma de entrada que contém unidades de processamento para cada elemento do vetor de entrada e a segunda é uma camada de saída da rede. As unidades de entrada estão ligadas com as de saída e o número de unidades de processamento na camada de saída é determinado pelo utilizador, com base na forma e tamanho inicial do mapa desejado. Ao contrário de outras redes neurais não há nenhuma camada oculta ou unidades de processamento escondidas. A figura a seguir apresenta um exemplo de uma rede SOM.



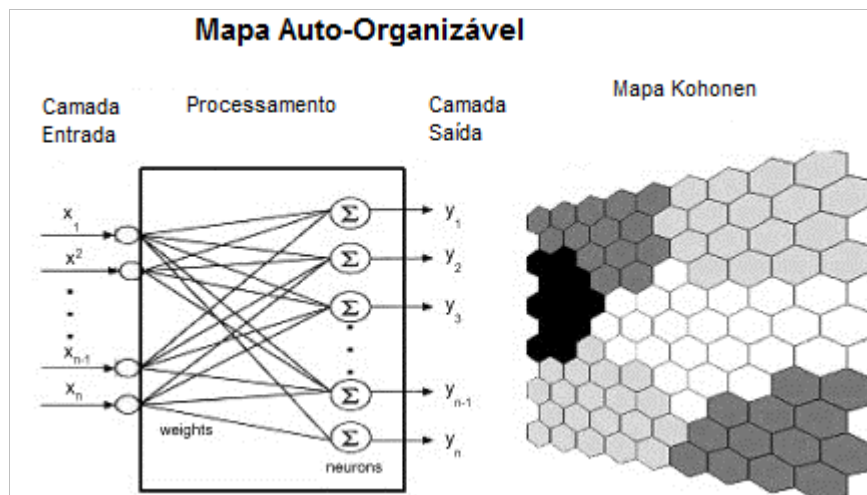


Figura 2.1 – Mapa Auto-Organizável

Quando um elemento de entrada é apresentado à rede, as unidades da camada de saída competem entre si. A unidade vencedora será àquela cujos pesos da conexão de entrada são os mais próximos do padrão de entrada em relação à distância Euclidiana. Assim, a entrada é apresentada e cada unidade de saída compete para coincidir com o padrão de entrada. A saída que está mais próxima do padrão de entrada é a vencedora, esta unidade é conhecida como *Best Matching Unit* (BMU). Os pesos de conexão da unidade vencedora são então ajustados, isto é, deslocados na direção do padrão de entrada através de um fator determinado taxa de aprendizagem (Deboeck & Kohonen, 1998).

Quanto ao tamanho do mapa, Ultsch & Moerchen (2005) acreditam que o poder de auto-organização que permite o surgimento de estruturas de dados é muitas vezes negligenciado pela utilização de mapas pequenos. Utilizar um mapa auto-organizável pequeno é quase idêntico a um *K-means*<sup>1</sup>, onde o número de neurónios é o mesmo que o número de *clusters* necessário. A preservação da topologia na projeção do SOM é de pouca utilidade quando se usa mapas pequenos. Fenómenos emergentes<sup>2</sup> envolvem um grande número de neurónios, onde grande significa pelo menos alguns milhares de neurónios. O uso de mapas grandes é conhecido como Mapas Auto-organizáveis Emergentes (*Emergent Self-Organizing Maps* - ESOM).

<sup>1</sup> **K-means** é um método de análise de cluster, que visa partição de  $n$  observações em  $k$  clusters em que cada observação pertence ao cluster com média mais próxima.

<sup>2</sup> **Fenómenos emergentes** geralmente estão associados a sistemas complexos e podem ser vistos como fenômenos que apresentam imprevisibilidade, surpresa na sua aparição.

A avaliação da qualidade do SOM é muito importante na produção da análise fiável. Existem dois tipos principais de erro no SOM: a qualidade da adaptação do SOM e os seus erros topológicos (Kohonen, 2001). As diferenças entre a posição de um dado padrão de entrada e a unidade onde se encontra mapeado é o erro de quantização e representa uma medida da qualidade da adaptação do SOM. Já os erros topológicos, por outro lado, estão preocupados com a preservação da topologia do SOM. O erro topológico mede o número médio de vezes em que o segundo vizinho mais próximo de um padrão de entrada não é mapeado para uma unidade vizinha da unidade onde o padrão de referência de encontra mapeado.

Quanto à visualização dos resultados, quando o número de neurónios é muito grande, como é o caso dos mapas auto-organizáveis emergentes, os *clusters* são detetados a partir da densidade das unidades. Para analisar a densidade do espaço de entrada no espaço de saída, a melhor maneira é através do uso de U-Matrizes (Ultsch & Siemon, 1990). Isto porque as U-Matrizes apresentam a distribuição das unidades numa paisagem 3D do espaço de dados de alta dimensionalidade.

No âmbito deste trabalho pretende-se utilizar as variáveis provenientes da Análise Técnica com mapas-auto organizáveis emergentes e consequentemente fazer uso das U-Matrizes (*Unified Distance Matrix*) para análise dos resultados. Mais detalhes sobre a metodologia será apresentada no próximo capítulo (3. Metodologia).

Neste capítulo foi apresentado um breve resumo dos mapas auto-organizáveis pois no âmbito do trabalho pretende-se utilizar o SOM como ferramenta para testar variáveis, não sendo objetivo do mesmo fazer testes ao algoritmo. Para mais detalhes sobre mapas auto-organizáveis consulte (Kohonen, 1982), (Kohonen T. , 2001), (Vesanto, 1999).

### **2.3. TRABALHOS RELACIONADOS**

Há uma larga e rica literatura sobre modelos de previsão de ativos financeiros, sendo que os principais objetivos geralmente são reproduzir o preço dos ativos financeiros cotados e fazer projeções de preços (Deboeck & Kohonen, 1998).

Até o momento não foram encontrados trabalhos de investigação cujo foco incidisse em indicadores da Análise Técnica e mapas auto-organizáveis com o objetivo de analisar as variáveis. Foram encontrados trabalhos que recaiam sob o tema dos mapas auto-organizáveis e sua utilização como auxiliar para um posterior algoritmo de

redes neurais supervisionado com o objetivo de previsão dos preços. Ou ainda para segmentação, mas que receberam como entrada variáveis da Análise Fundamental.

As ferramentas existentes para previsão de preços e tendências, que geralmente utilizam algoritmos como Redes Neurais, baseiam-se com frequência na Análise Técnica e utilizam indicadores como o histórico de preços (preço de abertura, preço mínimo, preço máximo e preço de fechamento), o volume de vendas, médias móveis, entre outros.

Por outro lado, as análises realizadas com mapas auto-organizáveis para o mercado financeiro utilizam normalmente variáveis de Análise Fundamental, como rentabilidade, estrutura de capital, custos financeiros, estrutura de risco, patrimônio líquido, entre outras, e possuem como principais objetivos a redução de dimensionalidade, visualização e a segmentação de diferentes componentes como bancos, ativos financeiros, fundos mútuos, entre outros.

Em seu livro *“Visual explorations in finance with self-organizing maps”* Deboeck & Kohonen (1998) apresentaram vários estudos realizados com os mapas auto-organizáveis na área financeira. A seguir apresenta-se resumidamente alguns dos estudos deste livro e também outros estudos relacionados com Análise Técnica e/ou mapas auto-organizáveis.

Kiviluoto & Bergius (1997) utilizaram os mapas auto-organizáveis como ferramenta para analisar demonstrações financeiras com foco na previsão de falências. O fenômeno da falência foi analisado qualitativamente em empresas de pequeno e médio porte, sendo as empresas classificadas em sadias ou pré-falidas. O estudo consistiu de duas partes: análise qualitativa e classificação. Ambas as partes utilizaram mapas auto-organizáveis. Na análise qualitativa, o SOM foi utilizado para formar uma regressão não-linear a partir do espaço de entrada para um plano, o que possibilitou examinar visualmente as diferenças entre as empresas que vão à falência e àquelas que não irão. Na classificação, o SOM foi utilizado como um vetor de quantização. Os indicadores financeiros escolhidos para treinar o SOM na parte da análise qualitativa foram quatro indicadores financeiros (não citados no trabalho) que medem a rentabilidade e a solvência de uma empresa.

Resta (1998) demonstrou experimentalmente o grande potencial de utilizar redes neurais para projetar um sistema para o mercado de ações. Ela propôs a utilização de uma arquitetura híbrida que combinou mapas auto-organizáveis com algoritmos genéticos, o que a autora chamou de *Integrated Self-Organization and*

*Genetic* (ISOG). Para o estudo utilizou a flutuação diária do índice *Milan Stock Exchange Official Index* e mostrou que a capacidade de previsão do sistema proposto foi superior aos resultados obtidos com o SOM.

E ainda em outras pesquisas mais recentes foram encontrados trabalhos como a pesquisa de Su & Huang (2003), no qual utilizaram indicadores da Análise Técnica como variáveis de entrada para prever a tendência de preços. No entanto, os indicadores técnicos não foram utilizados diretamente para determinar a tendência de ações, em vez disso, propuseram a utilização de um sistema *fuzzy* para transformar os indicadores técnicos em variáveis mais correlacionadas com a tendência de preço das ações. Recorreram ainda aos mapas auto-organizáveis para mensurar o grau em que os dados eram não-determinísticos. Em seguida modelaram uma rede neuronal para prever as tendências dos preços. Neste trabalho, utilizaram 5 indicadores técnicos: média móvel (MA), linha estocástica (KD), média móvel de convergência e divergência (MACD), índice de força relativa (RSI) e média móvel exponencial (EMA).

Li & Kuo (2008) propuseram um modelo de previsão híbrido, utilizando o histórico de preços do índice *Weighted Stock Index* (TAIEX) entre 1991 e 2002. Utilizaram Transformada Wavelet Discreta para se decompor uma série temporal em subsequências com diferentes escalas e para extrair características temporais ocultas. Empregaram ainda um mapa auto-organizável modificado com dois níveis para analisar a trajetória da série e facilitar a descoberta de conhecimento no processo de previsão.

Já Khan, Bandopadhyaya & Sharma (2008) utilizaram mapas auto-organizáveis em conjunto com redes neurais com retropropagação baseados em algoritmos genéticos para propor um método melhorado de seleção de ações. As variáveis utilizadas são provenientes da Análise Fundamental como o crescimento anual das vendas, crescimento anual do lucro apurado após impostos, retorno sobre o capital empregado, lucro líquido entre outras.

Mais recentemente, Atsalakis & Valavanis (2009) desenvolveram um sistema neuro-fuzzy composto por um Sistema de Inferência *Neuro Adaptive Fuzzy* (ANFIS) com o objetivo de prever as tendências do mercado de ações e, em particular, a tendência do dia seguinte das ações escolhidas. As variáveis utilizadas para treinar e avaliar o sistema proposto foram quinze diferentes combinações de preços do histórico das ações Atenas e *New York Stock Exchange* (NYSE).

Conforme apresentado nos trabalhos relacionados, o estudo de indicadores da Análise Técnica e mapas auto-organizáveis com o objetivo de analisar as variáveis ainda não foi explorado, por isso o desafio deste trabalho será analisar variáveis da Análise Técnica que serão baseadas no histórico de preços do índice *Dow Jones Industrial Average* com auxílio dos mapas auto-organizáveis. Os agrupamentos realizados pelo algoritmo não serão em relação a diferentes índices, mas a diferentes momentos no tempo. Pretende-se descobrir quais variáveis são capazes de descrever o comportamento do mercado e possivelmente sinalizar momentos de estabilidade e de crise.

### **3. METODOLOGIA**

Nesta secção é apresentada a metodologia utilizada para a realização deste trabalho, incluindo a classificação da pesquisa, fonte de dados, seleção das variáveis que serão utilizadas e modelo para processamento e análise de dados.

#### **3.1. CLASSIFICAÇÃO DA PESQUISA**

Esta pesquisa apresenta-se sob a perspectiva do paradigma positivista, uma vez que se parte da hipótese de que a utilização de variáveis de Análise Técnica na segmentação produzirá bons resultados na análise do mercado financeiro.

A abordagem será quantitativa, dado que se baseia em modelos numéricos, recorrendo a fórmulas e teoria matemática e estatística para explicar o comportamento económico.

A estratégia utilizada será experimental com horizonte longitudinal, pois o experimento será realizado com base em dados históricos do comportamento dos ativos.

O método de recolha de dados será a utilização de dados secundários, já que toda a informação necessária é publicada diariamente pelas bolsas de valores, sendo a informação confiável e pública.

#### **3.2. FONTE DE DADOS**

Para a realização do experimento será selecionado o índice *Dow Jones Industrial Average* que é um dos principais índices do mercado dos Estados Unidos, composto pela média ponderada dos preços das 30 empresas mais importantes do setor industrial.

O horizonte a estudar serão os últimos 10 anos, ou seja, o histórico dos preços dos índices inicia-se em 01 de Janeiro de 2001 e termina em 31 de Dezembro de 2011. O conjunto de dados é composto por 2767 observações.

A escolha deste período deve-se ao fato de se verificar grande variabilidade no comportamento dos mercados neste período, incluindo duas crises financeiras, nomeadamente em 2002 e 2008. A crise de 2002, conhecida como bolha da internet, marcou a queda acentuada nos preços das ações nas bolsas de valores nos Estados Unidos, Canadá, Ásia e Europa, com declínios mais acentuados em Julho e Setembro de 2002.

Já a crise de 2008, conhecida por Crise Financeira Global, foi segundo Greenspan, Mankiw & Stein (2010) provocada pela pesada securitização do mercado de hipotecas *subprime* dos EUA entre 2003 e 2006 que gerou o que o autor denominou de “ativos tóxicos” e o consequente colapso global em 2008. Esta crise é considerada por vários economistas como a pior crise financeira desde a grande depressão de 1929.

Sobre este histórico de preços serão calculadas as variáveis utilizadas na pesquisa, as quais serão explicadas a seguir.

A fonte de dados históricos utilizada será a Bloomberg. A Bloomberg é um dos principais provedores mundiais de informação para o mercado financeiro. Os terminais de informações Bloomberg estão presentes em quase 100% dos bancos, corretoras e seguradoras a nível mundial. Possuindo também uma emissora de TV a cabo, a Bloomberg Television que faz a transmissão ao vivo, das principais bolsas de valores ao redor do mundo, bem como entrevistas e matérias sobre o mercado financeiro. A Bloomberg publica também livros e revistas com conteúdo financeiro e relatórios diversos.

### 3.3. VARIÁVEIS

Nesta pesquisa foram selecionadas variáveis designadas como pertencentes à Análise Técnica. As variáveis selecionadas foram:

- ✓ **Médias móveis:** A média móvel simples (SMA) é uma técnica muito utilizada por analistas técnicos por ser o tipo de média móvel mais fácil de calcular, e também por ser bastante eficaz. Este tipo de média é equivalente a uma média aritmética móvel (Meyers, 2002). O principal objetivo da média móvel é fornecer o valor médio da cotação dentro de um determinado período (horas, dias, semanas, etc.). Assim, para cada valor incluído no cálculo da média, o valor mais antigo é excluído. Na média móvel simples, cada elemento utilizado no cálculo da média terá o mesmo peso, ou seja, a média móvel simples é calculada adicionando-se os preços para um determinado número de períodos, que nesta pesquisa corresponderá a dias, e dividindo-se esse valor pelo número de períodos. A seguir apresenta-se a fórmula para o cálculo da média móvel simples:

$$SMA(t)n = \frac{P(t-1) + P(t-2) + P(t-n)}{n}$$

Onde:

$SMA(t)n$  = Média móvel simples de um período (n) num dado tempo (t)

$P(t-1)$  = Preço no tempo (t-1)

n = Número de períodos (dias, semanas, etc.)

Serão utilizadas médias para representar os períodos de curto médio e longo prazo, ou seja, serão utilizadas médias móveis de 10, 20, 30, 40, 50, 60, 100, 120 e 180 dias com o objetivo de analisar qual o intervalo mais significativo para análise do movimento dos preços. O preço utilizado para o cálculo desta variável será o preço mínimo por representar o momento mais pessimista da flutuação diária.

- ✓ **Volume:** refere-se ao número de ativos financeiros transacionados num dado período, como o número de ações, contratos futuros, ou mesmo o valor em termos monetários, como exemplo a quantidade de dinheiro transacionada num certo período de tempo no mercado cambial, entre outros (Ferreira, 2011). No âmbito deste trabalho o volume será o número de ações negociadas no período de 1 dia. O volume é um dos indicadores base fornecidos pela base de dados da Bloomberg.

A análise do volume é um elemento básico, contudo muito importante da Análise Técnica. Ela fornece indicadores a respeito da intensidade de movimentos relacionados com o preço. Além disso, o volume pode confirmar a força de uma tendência. Volumes altos são comuns onde há um forte consenso de que os preços irão aumentar. Volumes baixos geralmente ocorrem quando um ativo se está a encaminhar para uma tendência de queda. Eles são caracterizados por expectativas indefinidas que acontecem em períodos de consolidação.

- ✓ **Volume Médio:** No caso do volume é importante analisar não apenas o seu valor integral mas também a sua variação para perceber se os



movimentos dos preços são sustentados pelo aumento do volume de negócios. O volume médio pode ser representado pela seguinte equação:

$$VM(t)n = \frac{V(t-1) + V(t-2) + V(t-n)}{n}$$

Onde:

$VM(t)n$  = Volume médio de um período (n) num dado tempo (t)

$V(t-1)$  = Volume no tempo (t-1)

n = Número de períodos (dias, semanas, etc.)

Para facilitar a leitura desta variável dividiu-se o volume diário pelo volume médio dos últimos  $n$  dias. Assim, o resultado é um número relativo que indica o comportamento do volume diário em relação à sua média. Se o resultado for menor que 1 significa que o volume diário é menor do que a média, ao contrário se for maior que 1 indica que o volume diário superou o valor da média. Portanto, o cálculo do novo coeficiente de volume médio  $CVM(t)$  está representado a seguir:

$$CVM(t) = \frac{V(t)}{VM(t)}$$

Serão calculados os coeficientes de volumes médios  $CVM(t)$  para 5, 8, 10, 15, 16 e 20 dias, pois são considerados intervalos de curto prazo. Esta variável será calculada em intervalos de curto prazo pois são as mais usuais nos indicadores de tendência da Análise Técnica.

- ✓ **Preço:** existem quatro tipos de preços que estão associados a um ativo financeiro no mercado de ações:
  - Abertura: é o valor da primeira transação efetuada num dado período de tempo.
  - Máximo: é o preço mais alto a que um ativo foi transacionado num determinado período de tempo.
  - Mínimo: é o preço mais baixo a que um ativo foi transacionado em certo período de tempo.

- Fechamento: é o último preço a que um ativo foi transacionado no período (Ferreira, 2011).

O preço está sempre relacionado com o período de tempo. Nesta pesquisa será utilizado o período em dias. Os indicadores de preço são fornecidos pela base de dados da Bloomberg.

- ✓ **Índice de Hurst:** O valor desse expoente varia entre 0 e 1. Para  $H = 0.5$  o sinal ou processo é aleatório. Para  $0 < H < 0.5$ , o sinal é caracterizado como anti-persistente, ou seja, existe uma probabilidade maior do que cinquenta por cento de que um valor “negativo” seja seguido de um valor “positivo”. E para  $0.5 < H < 1$ , o sinal é dito persistente, pois apresenta uma tendência, isto é, a probabilidade de repetição de um valor é maior do que cinquenta por cento. O cálculo do índice de Hurst é apresentado nos Anexos, capítulo 7.1 Cálculo do Índice de Hurst. Os intervalos a serem testados neste trabalho serão 30, 40, 100 e 180 dias.

- ✓ **Variação Preço:** A variação do preço será uma variável calculada com base nas variáveis do preço e média móvel. Esta variável será representada por um número relativo  $[0, \infty]$  que mostra o valor preço atual em relação ao passado, ou seja, revela quando o valor atual de um ativo está subvalorizado, com valor médio ou supervalorizado em relação ao histórico.

A variação do preço  $VP(t)_n$  será calculada pela divisão do preço diário  $P(t)$  pelas médias móveis  $SMA(t)_n$ , conforme a equação abaixo:

$$VP(t)_n = \frac{P(t)}{SMA(t)_n}$$

A variação do preço será calculada para períodos de 10, 20, 30, 60, 100, 120 e 180 dias. Além das variações de curto, médio e longo prazo também será incluída a variação do preço em relação ao dia anterior para representar a flutuação diária.

- ✓ **Variação Preço Médio:** A variação do preço médio será uma variável calculada com base na média móvel. Esta variável será calculada pela divisão da média móvel  $SMA(t)n$  pela mesma média móvel do dia anterior  $SMA(t - 1)n$  e representa a flutuação diária das médias móveis.

$$VPM(t)n = \frac{SMA(t)n}{SMA(t - 1)n}$$

As médias móveis utilizadas para o cálculo da variação do preço médio serão de 10, 20, 30, 60, 100, 120 e 180 dias.

A seguir apresenta-se uma tabela com o resumo das variáveis utilizadas na pesquisa.

<b>Variável</b>	<b>Descrição</b>
px_open	Preço de Abertura
px_high	Preço Máximo
px_low	Preço Mínimo
px_close	Preço de Fechamento
px_volume	Número de ações negociadas em 1 dia
ix_hurst_30	Índice de Hurst calculado com intervalos de 30 dias
ix_hurst_40	Índice de Hurst calculado com intervalos de 40 dias
mov_avg_10d	Média móvel 10 dias
mov_avg_20d	Média móvel 20 dias
mov_avg_30d	Média móvel 30 dias
mov_avg_40d	Média móvel 40 dias
mov_avg_50d	Média móvel 50 dias
mov_avg_60d	Média móvel 60 dias
mov_avg_100d	Média móvel 100 dias
mov_avg_120d	Média móvel 120 dias
mov_avg_180d	Média móvel 180 dias
var_preco_1	Variação do Preço 1 dia
var_preco_10	Variação do Preço 10 dias
var_preco_20	Variação do Preço 20 dias
var_preco_30	Variação do Preço 30 dias
var_preco_60	Variação do Preço 60 dias
var_preco_100	Variação do Preço 100 dias
var_preco_120	Variação do Preço 120 dias
var_preco_180	Variação do Preço 180 dias
volume_medio_5	Volume médio 5 dias
volume_medio_8	Volume médio 8 dias
volume_medio_10	Volume médio 10 dias
volume_medio_15	Volume médio 15 dias
volume_medio_16	Volume médio 16 dias
volume_medio_20	Volume médio 20 dias
var_preco_medio_10	Variação do Preço Médio 10 dias
var_preco_medio_20	Variação do Preço Médio 20 dias
var_preco_medio_30	Variação do Preço Médio 30 dias
var_preco_medio_60	Variação do Preço Médio 60 dias
var_preco_medio_100	Variação do Preço Médio 100 dias
var_preco_medio_120	Variação do Preço Médio 120 dias
var_preco_medio_180	Variação do Preço Médio 180 dias

Tabela 3.1 - Tabela das Variáveis

### 3.4. REDUÇÃO DAS VARIÁVEIS

Inicialmente partiu-se de um conjunto com 37 variáveis que compõem 7 diferentes medidas: médias móveis, volume, preço, índice de Hurst, volume médio, variação do preço e variação do preço médio. Entretanto, como se trata de uma amostra pequena, composta por 2767 observações, torna-se necessário reduzir o conjunto de variáveis para evitar a maldição da dimensionalidade.

Maldição da dimensionalidade é o nome conhecido do fenómeno que ocorre quando há alta dimensionalidade no espaço de entrada (ou seja, alto número de atributos), o que provoca também o aumento do tamanho do espaço de pesquisa de forma exponencial e assim os dados ficam mais esparsos, tornando-se difícil encontrar grupos de pontos semelhantes. Isto pode prejudicar a construção de modelos (Maimon & Rokach, 2005).

À medida que a dimensão do espaço de entrada cresce, requer-se também maior quantidade de exemplos de treino. Portanto, como o número de exemplos deste trabalho é pequeno, é importante reduzir o número de variáveis redundantes para não prejudicar a eficiência do modelo. Para este fim, foi utilizado o procedimento VARCLUS procedure (SAS/STAT®).

O procedimento PROC VARCLUS é considerado um bom método de redução de variáveis. Este procedimento divide um conjunto de variáveis numéricas em *clusters* disjuntos ou hierárquicos. Associado a cada *cluster* está uma combinação linear das variáveis do *cluster*, a qual pode ser a componente principal ou o componente centróide. A combinação linear utilizada neste trabalho foi a componente principal. A componente principal é uma média ponderada das variáveis e que explica a variância, tanto quanto possível (SAS/STAT® User's Guide).

O procedimento PROC VARCLUS tenta maximizar a somatória entre *clusters* de variância das variáveis originais que são explicadas pelos *clusters* de componentes. Cada correlação da matriz de covariância é analisada. Se a correlação for utilizada como parâmetro para o procedimento, todas as variáveis são tratadas como igualmente importantes, caso contrário, se a covariância for utilizada, as variáveis com maior variância têm maior importância na análise (SAS/STAT® User's Guide). Neste trabalho o parâmetro utilizado será a correlação, pois interessa-nos encontrar as variáveis que estão correlacionadas entre si com o objetivo de eliminá-las do conjunto de entrada.

Os testes realizados com as variáveis são apresentados no capítulo 4 Resultados e Discussão.

### 3.5. ANÁLISE DE SUBCONJUNTOS

Inicialmente foi selecionado um conjunto de 10 anos desde 2001 a 2011. E, entretanto, com as análises realizadas por Silva, Panosso & Marques (2012) com auxílio da metodologia *Concept Drift*, identificou-se que um período que sugere estabilidade do mercado corresponde ao período que inicia após a recuperação da queda de 2002 até ao período que antecede a outra grande queda em 2008, aproximadamente entre Julho de 2002 e Junho de 2007.

Este subconjunto é particularmente interessante no âmbito deste trabalho pois permitir-nos-á estudar as características de uma situação de estabilidade onde genericamente os investidores podem investir tranquilamente no mercado de ações, pois nestes períodos a tendência de longo prazo aponta para aumento dos preços.

Segue a visualização no gráfico representando os preços dos períodos considerados neste trabalho, nomeadamente o período de 10 anos de Janeiro-2001 a Dezembro-2011 e o período Julho-2002 a Junho-2007.

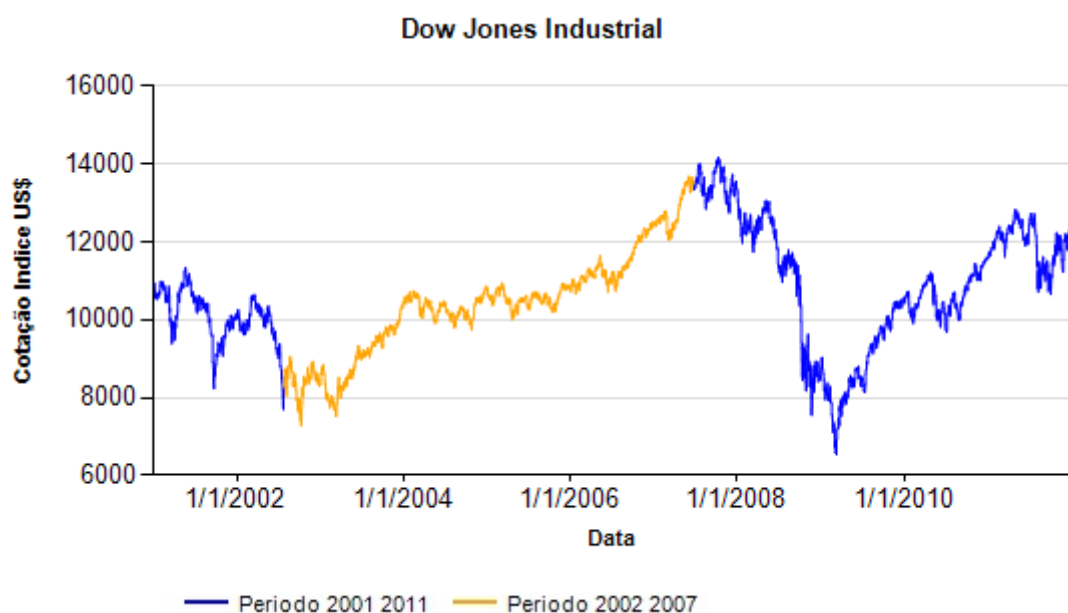


Figura 3.1 - Períodos de estudo

### 3.6. PROCESSAMENTO E ANÁLISE DE DADOS

A técnica utilizada para processamento e análise dos dados deste trabalho será o Mapa Auto-Organizável (*Self-Organizing Map* - SOM), também conhecido como mapas de Kohonen.

Conforme apresentado no capítulo 2.2, o SOM é uma rede neuronal que utiliza um algoritmo de treinamento não supervisionado, e através de um processo chamado de auto-organização configura as unidades de saída em uma representação topológica dos dados originais. O SOM pode ser visto como uma matriz de unidades (neurónios) e o conjunto de variáveis são as entradas da rede. Associado a cada variável está um determinado peso que vai sendo ajustado durante o treino do SOM.

Antes de iniciar o treino do SOM, é preciso determinar os parâmetros utilizados na fase de treino, ou seja, a topologia do mapa. O primeiro parâmetro a ser definido é o tamanho do mapa, isto é, a quantidade de neurónios que serão utilizados na matriz. A ferramenta Som Toolbox, que será utilizada no desenvolvimento deste trabalho, utiliza a fórmula a seguir para determinar o tamanho do mapa (Alhoniemi, Himberg, Parhakangas & Vesanto, 2012).

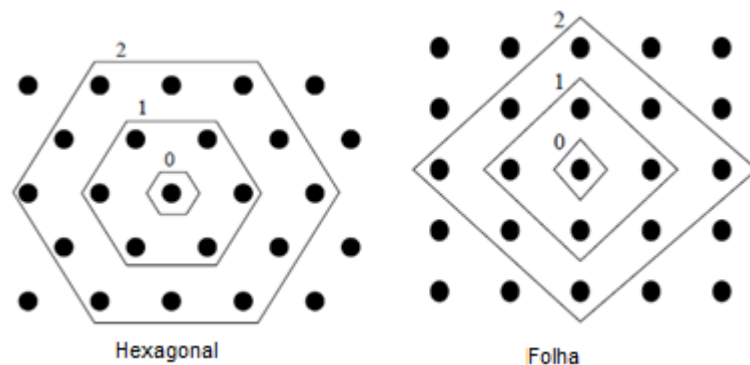
Tamanho do Mapa	Fórmula
Normal	$5 * \sqrt{dlen}^3$
Pequeno	$\frac{5 * \sqrt{dlen}}{4}$
Grande	$4 * (5 * \sqrt{dlen})$

Tabela 3.2 - Fórmula tamanho do mapa

Outro parâmetro que determina a topologia é a disposição dos neurónios na matriz, a qual definirá a vizinhança de cada neurónio. A disposição do neurónio pode ser retangular ou hexagonal, tal como apresentado na Figura 3.2.

---

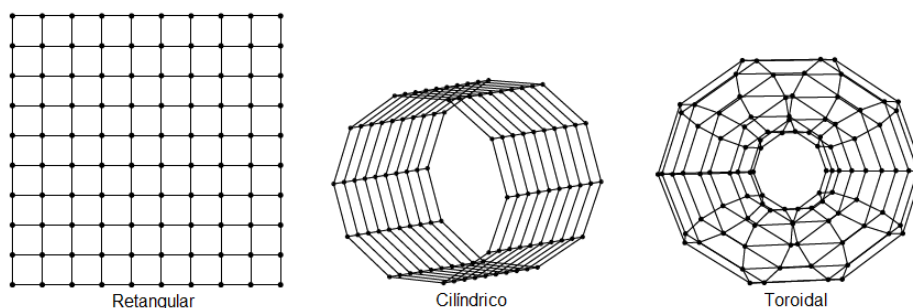
<sup>3</sup> **Dlen** representa o número de observações da amostra



Fonte: (Vesanto, Himberg, Alhoniemi & Parhankangas, 2000) (Adaptada)

Figura 3.2 - Topologia de vizinhança

Quanto à forma da matriz, o SOM pode assumir 3 variações: Retângulo, Cilindro ou Toróide. O formato padrão utilizado para visualização do mapa é a forma retangular. Se o objetivo é obter uma visualização bidimensional fechada, utiliza-se o formato cilíndrico ou caso se opte por uma visualização tridimensional é sugerido o formato toroidal.



Fonte: (Vesanto, Himberg, Alhoniemi & Parhankangas, 2000) (Adaptada)

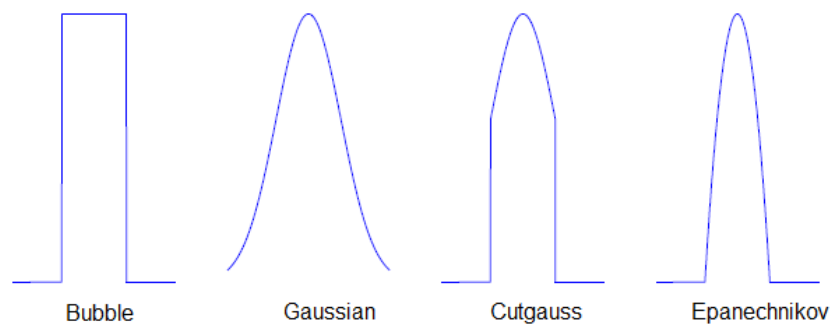
Figura 3.3 - Formas da matriz

Para a função vizinhança, existem quatro versões possíveis: Bubble, Gaussian, Cutgauss, Epanechnikov:

- **Bubble:** Define uma largura ou um raio do neurónio vencedor, e somente os neurónios que estão no alcance deste raio são ajustados ao padrão de entrada.
- **Gaussian:** É a mais usada. Uma importante propriedade desta função é que um neurónio que se encontra mais próximo do padrão de entrada sofrerá maior ajuste do que um neurónio que esteja mais afastado.



- **Cutgauss:** Esta é uma combinação das duas versões acima. Se a distância de um dado neurónio ao neurónio vencedor estiver dentro de um valor dado (raio), o neurónio será atualizado.
- **Epanechnikov:** Uma propriedade dessa função é que a taxa de aprendizagem é reduzida enquanto a distância de cada neurónio ao padrão de entrada aumenta.



Fonte: (Vesanto, Himberg, Alhoniemi & Parhankangas, 2000) (Adaptada)

Figura 3.4 - Funções de vizinhança

Outra característica importante a ser definida nos parâmetros de treino do SOM é o tipo de normalização das variáveis. A utilização de funções de normalização é importante para colocar todas as variáveis numa mesma escala, pois o algoritmo utilizado no SOM usa a distância euclidiana entre vetores, assim se houver valores em escalas muito grandes e valores muito pequenos no conjunto de dados, certamente os valores muito grandes terão um maior impacto nas medidas das distâncias. Na ferramenta Som Toolbox existem seis tipos de normalizações que são usadas na fase de pré-processamento:

- 'var' – É o tipo de normalização Z-score. A variância é normalizada a um e a média a zero.
- 'range' – É o tipo de normalização Min-Max. Os valores são normalizados entre [0,1] (transformação linear).
- 'log' – Logaritmo natural é aplicado aos valores:  $\text{novox} = \log(x - m + 1)$  onde  $m = \min(x)$ .
- 'logistic' – Método que escala todos os valores possíveis entre [0,1].
- 'histD' – Equalização discreta do histograma. Escala os valores linearmente [0,1].

- 'histC' – Equalização contínua do histograma. É semelhante ao anterior, porém com equalizações aproximadas. Os valores são escalados entre [0,1].

Por fim, é preciso definir a duração do treino. O número de iterações no SOM define quanto tempo a fase de treinamento durará. Assim, o número de iterações define o número de padrões de entrada apresentados à rede. Na ferramenta GeoSom Suite (ISEGI-UNL, 2013), há um parâmetro que é o número de épocas. Uma época é o conjunto de iterações em que todos os elementos do conjunto de entrada foram apresentados à rede. Esta apresentação dos padrões de entrada pode ser aleatória ou seguir a ordem do conjunto de dados. O grau de adaptação pode ser mensurado através medidas de qualidade do SOM, nomeadamente os erros de quantização e de topologia.

O processo de treino inicia-se após definidos os parâmetros iniciais. A seguir os pesos das unidades são gerados aleatoriamente sendo então apresentado à rede o primeiro elemento do conjunto de dados. Na sequência são avaliadas as distâncias entre o objeto e todas as unidades, sendo que a unidade mais próxima será a unidade escolhida para a representação do elemento. Uma vez definida a unidade mais próxima ao elemento, esta irá mover-se no sentido do elemento e arrastará consigo as unidades vizinhas, ocorrendo o mesmo sucessivamente a todos os elementos apresentados à rede. Dessa forma a rede adaptar-se-á aos elementos que representa. A aproximação que ocorre entre a unidade da rede e o elemento apresentado é definida pela taxa de aprendizagem, que varia entre 1 e 0. Quanto mais próxima de 1, mais rapidamente a rede mover-se-á.

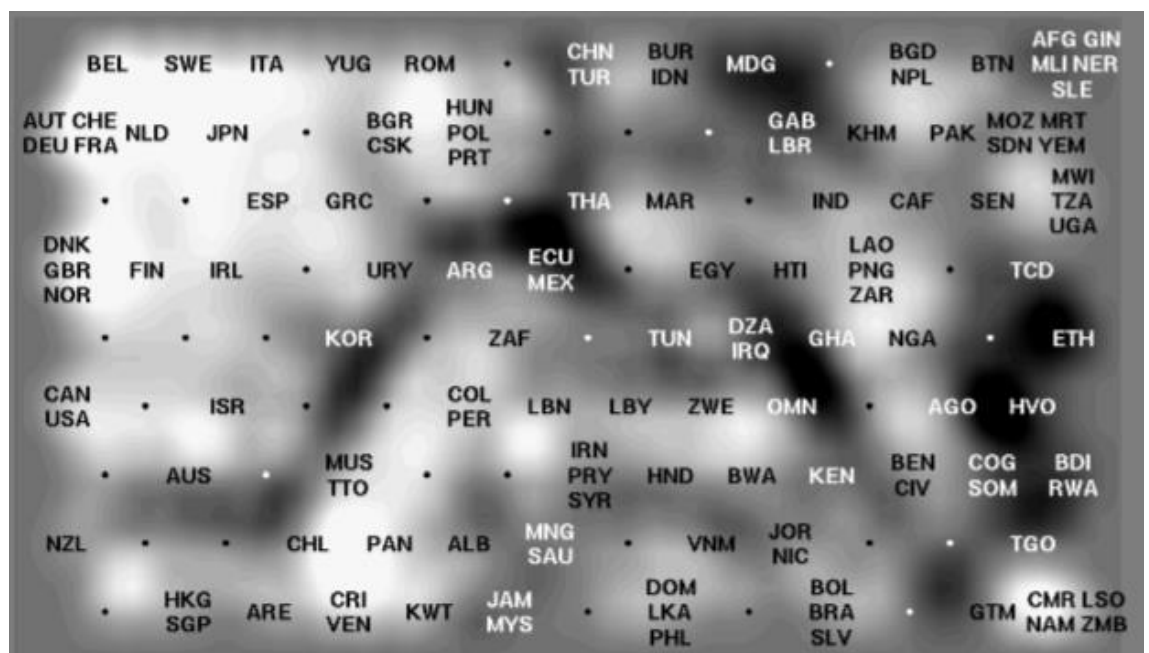
Neste trabalho os parâmetros iniciais da rede, ou seja, a topologia do SOM será definida através de testes aos diferentes tipos de parâmetros e análise inicial dos parâmetros mais adequados para o conjunto de dados do índice DJI.

Como resultado, no âmbito deste trabalho, a rede deverá agrupar os movimentos de preços do índice DJI. Isto permitirá analisar os diferentes segmentos (*clusters*) e identificar as características das variáveis que definem cada partição.

Os dados serão analisados através do uso das U-Matrizes. Segundo Ultsch & Siemon (1990), a U-matrix contém uma aproximação geométrica correta da distribuição do vetor na rede de Kohonen (SOM). Para obter uma impressão visual da distribuição dos neurónios pela rede, a U-matrix em três dimensões apresenta os

elementos com um valor de altura acima de uma grelha, criando uma paisagem 3D do espaço de dados de alta dimensionalidade. Esta altura do neurónio é grande em áreas onde há poucos dados ou nenhum, criando cadeias de montanhas para limites de *cluster*. A altura é pequena em áreas de alta densidade, assim, os *clusters* são retratados como vales.

A Figura 3.5 mostra um exemplo de U-Matrix, referente ao projeto *World Poverty Map*, que apresenta a qualidade de vida em 129 países de acordo com o seu nível de desenvolvimento. Neste caso a U-Matrix foi suavizada e não mostra as células que a formam. A ordenação dos países, que contém nomes abreviados, indica as semelhanças entre a qualidade de vida dos países. Pode-se observar, por exemplo, Portugal (código PRT) encontra-se na zona superior central da matriz, ligeiramente à esquerda, acompanhado pela Polónia (POL) e a Hungria (HUN), numa área que, apesar de próxima dos países mais desenvolvidos parece haver particularidades. As áreas mais claras representam as pequenas distâncias (os vales) onde há formação de *clusters*, já as cores mais escuras correspondem a grandes distâncias (as montanhas) o que determina separação entre *clusters*.



Fonte: (Kaski, Venna & Kohonen, 2000)

Figura 3.5 – U-Matrix *World Poverty Map*

A ferramenta GeoSOM Suite (ISEGI-UNL, 2013) que será utilizada neste trabalho permite que o utilizador crie seleções de clusters. Com base nas distâncias da U-Matrix, a ferramenta sugere uma segmentação de acordo com o número de *clusters* desejado. A figura 3.6 mostra o resultado de uma segmentação. Na imagem à esquerda da figura, os *clusters* estão representados por números, onde a formação de cluster é caracterizada pelo conjunto de neurónios com mesmo número. Na imagem ao centro apresenta-se uma seleção manual do *cluster* e na imagem à direita os segmentos foram desativados permitindo visualizar melhor as fronteiras do segmento. A ferramenta permite ainda exportar os resultados da segmentação manual. Este será o método de seleção de *clusters* utilizado neste trabalho.

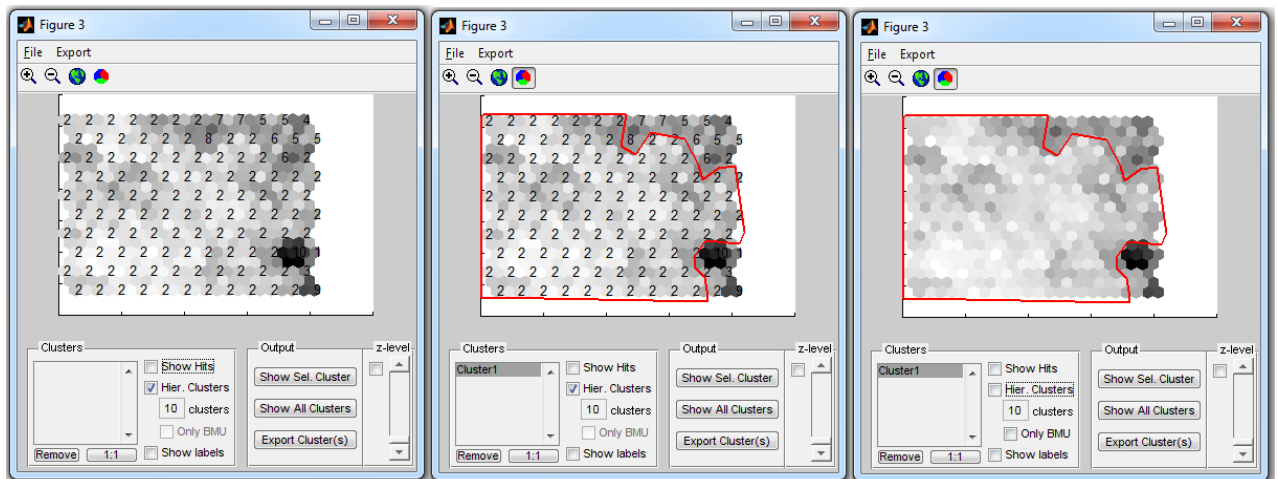


Figura 3.6 – Segmentação GeoSOM Suite

As ferramentas utilizadas para desenvolvimento deste estudo serão:

- GeoSOM Suite – treino e segmentação
- Matlab – Matriz de Labels
- Oracle – armazenamento do conjunto de dados e rotinas de cálculo das variáveis
- Microsoft Report Services – gráficos resultados
- SAS Enterprise Miner – *Cluster* das variáveis

## 4. RESULTADOS E DISCUSSÃO

### 4.1. REDUÇÃO DAS VARIÁVEIS

O índice de Hurst calculado a longo prazo não apresentou grande variação no valor do expoente para o conjunto de dados em análise. Portanto, os intervalos a serem testados neste trabalho serão 30 e 40 dias, já que apresentam maior amplitude e podem revelar momentos em que o mercado financeiro apresenta sinal aleatório ou anti-persistente.

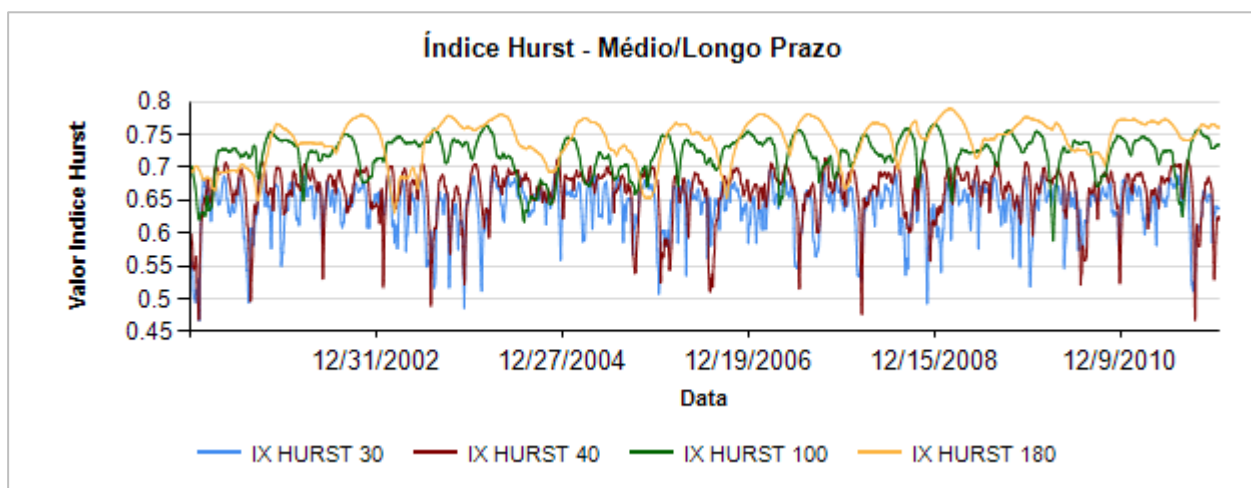


Figura 4.1 - Índice de Hurst Médio e Longo Prazo: Período 2001 a 2011

A seguir é apresentada a análise das variáveis utilizando o procedimento PROC VARCLUS. O PROC VARCLUS cria um conjunto de dados de saída que pode ser utilizado pelo procedimento TREE para desenhar um diagrama de árvore de agrupamentos hierárquicos (SAS/STAT® User's Guide). Esta árvore é ilustrada na figura abaixo:

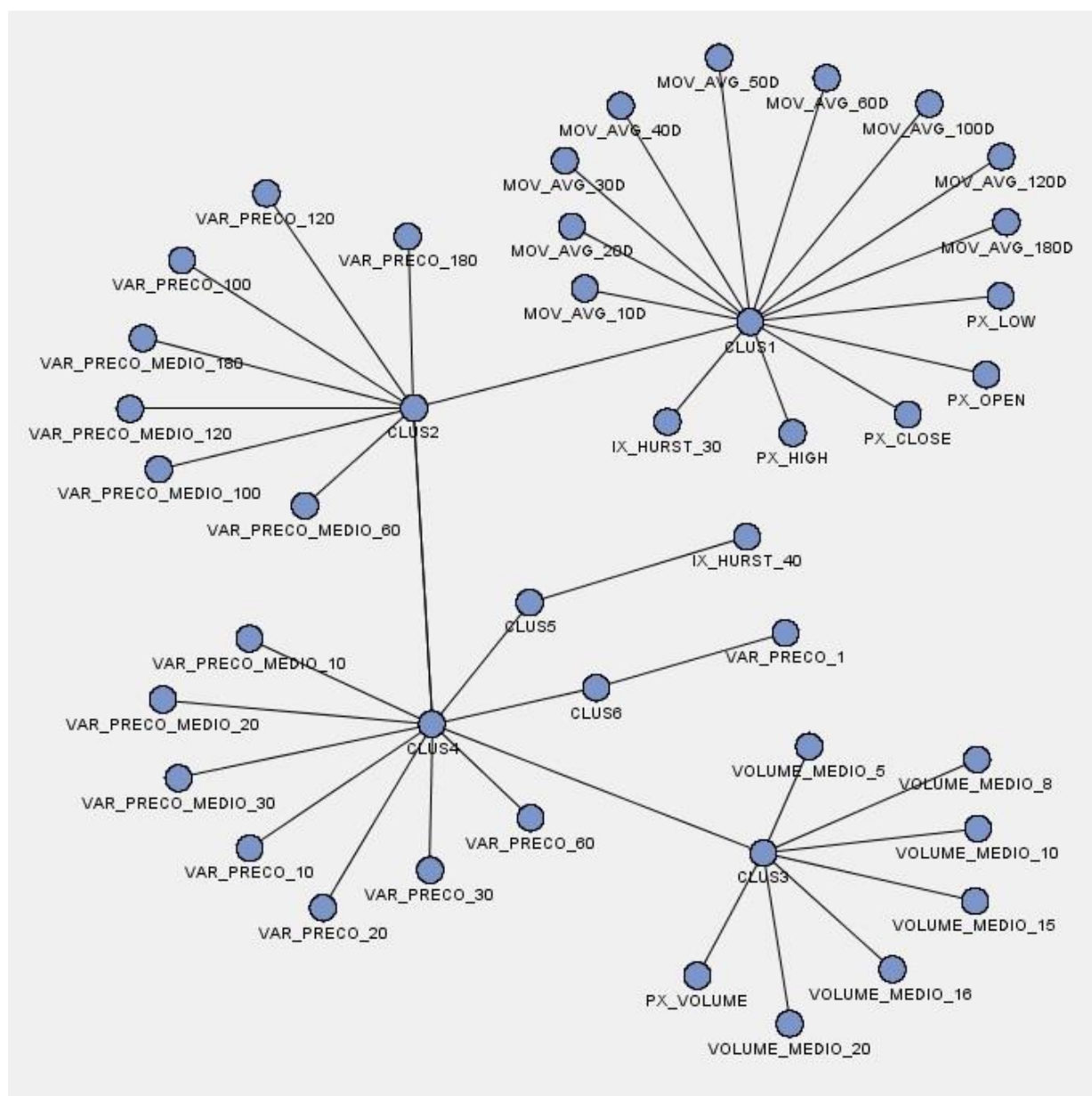


Figura 4.2 - Árvore clusters de variáveis

Como se pode ver na imagem acima, o procedimento PROC VARCLUS dividiu as 37 variáveis de entrada em 6 *clusters* de acordo com as correlações identificadas entre as variáveis. No *cluster* número 1 (CLUS1) aparecem as variáveis relacionadas com o conceito de preço (PX\_OPEN, PX\_HIGH, PX\_LOW, PX\_CLOSE) de médias móveis (MOV\_AVG\_10D, MOV\_AVG\_20D, etc.), apresentando todas elas alta correlação independentemente do intervalo de tempo que está representado, ou seja, uma média móvel de 30 dias (MOV\_AVG\_30D) apresenta correlação a uma média móvel de longo prazo (180 dias: MOV\_AVG\_180D). A correlação existente entre estas variáveis

faz muito sentido uma vez que as médias móveis são variáveis calculadas com base nos preços de um ativo. Existe ainda uma outra variável neste *cluster* que conceitualmente não apresenta correlação com as outras que é o índice de Hurst calculado a 30 dias (IX\_HURST\_30), embora seja calculado também com base no preço mínimo (PX\_LOW). Por esse motivo, esta variável será mantida no conjunto de variáveis de entrada, já em relação às restantes, serão realizados testes iniciais para decidir quais deverão ser mantidas neste subconjunto.

No CLUS2, encontram-se as variáveis variação do preço médio nos intervalos de médio e longo prazo (VAR\_PRECO\_MEDIO\_60, VAR\_PRECO\_MEDIO\_100, etc.) e variação do preço também de médio e longo prazo (VAR\_PRECO\_100, VAR\_PRECO\_120, VAR\_PRECO\_180). A presença deste subconjunto revela a importância de selecionar variáveis que representem os preços em diferentes períodos de tempo, isto é, variáveis de curto, médio e longo prazo. Além disso, a correlação entre as variáveis VAR\_PRECO e VAR\_PRECO\_MEDIO talvez venha a revelar que estes dois grupos de variáveis são coincidentes, não influenciando no resultado do modelo caso seja excluído um dos conceitos: variação do preço ou variação do preço médio.

No CLUS3 estão as variáveis relacionadas com o volume, nomeadamente o próprio volume (PX\_VOLUME) e volume médio (VOLUME\_MEDIO\_5, VOLUME\_MEDIO\_8, etc.). Neste caso, a variável volume será mantida e as médias serão avaliadas com experimento para determinar quais serão incluídas no conjunto final.

No CLUS4 aparecem novamente as variáveis variação do preço médio (VAR\_PRECO\_MEDIO) e variação do preço (VAR\_PRECO), mas agora com intervalos de curto e médio prazo. Salienta-se novamente a alta correlação entre estes dois conceitos de variáveis.

No CLUS5 surge somente a variável índice de Hurst calculada a 40 dias (IX\_HURST\_40), o que significa que esta variável não está correlacionada com nenhuma outra variável do conjunto, nem mesmo o índice de Hurst calculado a 30 dias. O mesmo ocorre no CLUS6, onde se verifica a presença da variável VAR\_PRECO\_1 que representa a flutuação diária de preço. Esta variável também deve ser incluída no conjunto final de entrada. No que diz respeito às variáveis que estão correlacionadas com outras, devem ser analisados os resultados dos testes iniciais para identificar quais são mais significativas e deverão permanecer no estudo.

Para definir o conjunto final de variáveis, foram realizados testes com diferentes conjuntos tendo por base as correlações apresentadas anteriormente e a análise dos erros de quantização dos diferentes conjuntos. Dessa forma, foi selecionado o conjunto que apresentou menor erro de quantização e também melhor segmentação dos dados. Convém no entanto destacar, a necessidade de efetuar uma análise cuidadosa ao avaliar os conjuntos, pois quanto menor o conjunto de variáveis menor o erro de quantização, perdendo-se no entanto na capacidade de separação dos dados. O conjunto final é composto por 12 variáveis:

- **PX\_VOLUME:** Número de ações negociadas em 1 dia
- **VOLUME\_MEDIO\_10:** Volume médio 10 dias
- **IX\_HURST\_30:** Índice de Hurst calculado com intervalos de 30 dias
- **IX\_HURST\_40:** Índice de Hurst calculado com intervalos de 40 dias
- **VAR\_PRECO\_1:** Variação do preço 1 dia que representa a flutuação diária do preço mínimo.
- **VAR\_PRECO\_10:** Variação do preço 10 dias
- **VAR\_PRECO\_20:** Variação do preço 20 dias
- **VAR\_PRECO\_30:** Variação do preço 30 dias
- **VAR\_PRECO\_60:** Variação do preço 60 dias
- **VAR\_PRECO\_100:** Variação do preço 100 dias
- **VAR\_PRECO\_120:** Variação do preço 120 dias
- **VAR\_PRECO\_180:** Variação do preço 180 dias

As variáveis de preço (PX\_OPEN, PX\_HIGH, PX\_LOW, PX\_CLOSE) e médias móveis (MOV\_AVG) que presentes no CLUS1, foram excluídas por não apresentarem ganhos ao modelo. Em oposição, apresentaram maiores erros de quantização, e além disso, comprometeram a segmentação pelo excesso de ordenação dos resultados. Este comportamento pode ser explicado pelo fato das variáveis de preço representarem o “target” do modelo, ou seja, se o objetivo do modelo é compreender e explicar o comportamento dos preços dos ativos financeiros, logo a variável a ser explicada não deverá fazer parte do conjunto de entrada.

Já as variáveis de variação do preço médio (VAR\_PRECO\_MEDIO) não se apresentaram significativas ao modelo e uma vez que estavam correlacionadas com as variáveis de variação do preço (VAR\_PRECO), optou-se pela sua exclusão. No caso das



variáveis de variação do preço (VAR\_PRECO), como estas demonstraram maior significância para os resultados do modelo, as mesmas foram incluídas no conjunto de entrada.

#### 4.2. TAMANHO DO MAPA

O tamanho de mapa utilizado inicialmente para análise das variáveis foi o tamanho determinado pela fórmula dada pela SOM Toolbox, para o tamanho Normal, ou seja, como a amostra contém 2767 observações, o tamanho Normal corresponde a  $5 * \sqrt{2767} = 263$  neurónios o que determina uma matriz de 18x15. Esta foi a matriz utilizada para analisar os erros dos diferentes conjuntos de variáveis descritos no item anterior.

Em seguida, foram testados vários tamanhos de matrizes variando o número de neurónios correspondente do tamanho pequeno até ao tamanho normal. Selecionou-se o maior tamanho com menor erro de quantização e topologia. Segue a análise do conjunto de 10 anos, composto por 2767 observações:

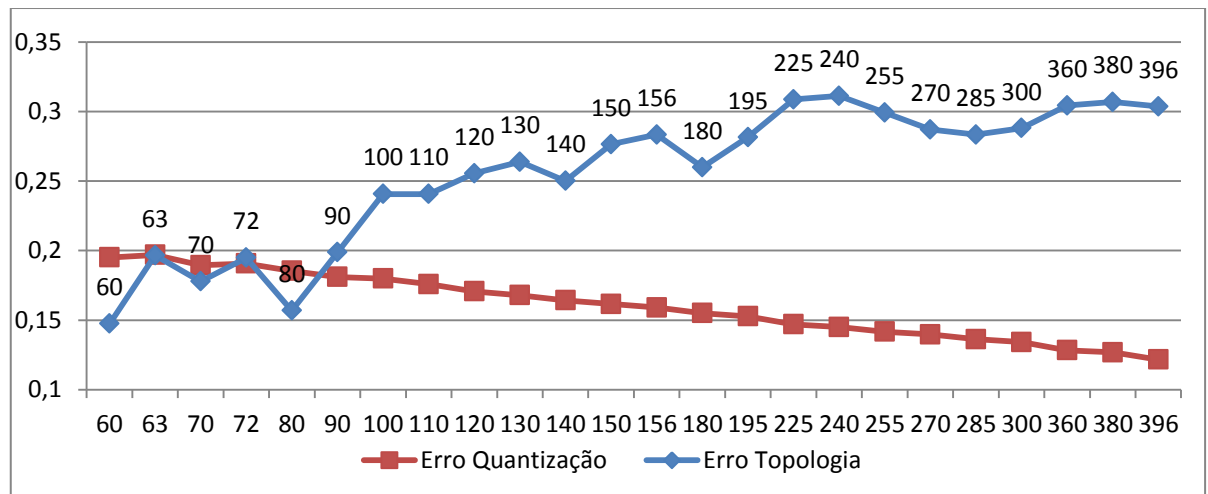


Figura 4.3 - Gráfico de Erros x Número de neurónios – 2001 a 2011

Neste caso, as melhores matrizes seriam as de 80, 140, 180 e 285 neurónios. Portanto, foi selecionada a maior matriz de 19x15 (285 neurónios) para se conseguir uma melhor visualização dos dados.

No caso do subconjunto compreendido entre 2002 e 2007, este é composto por 1236 observações, correspondendo aqui o tamanho Normal a 175 neurónios. Neste

caso, a matriz mais próxima deste número de neurónios e que apresentou menor erro de quantização e topologia foi a matriz 15x13, ou seja, 195 neurónios.

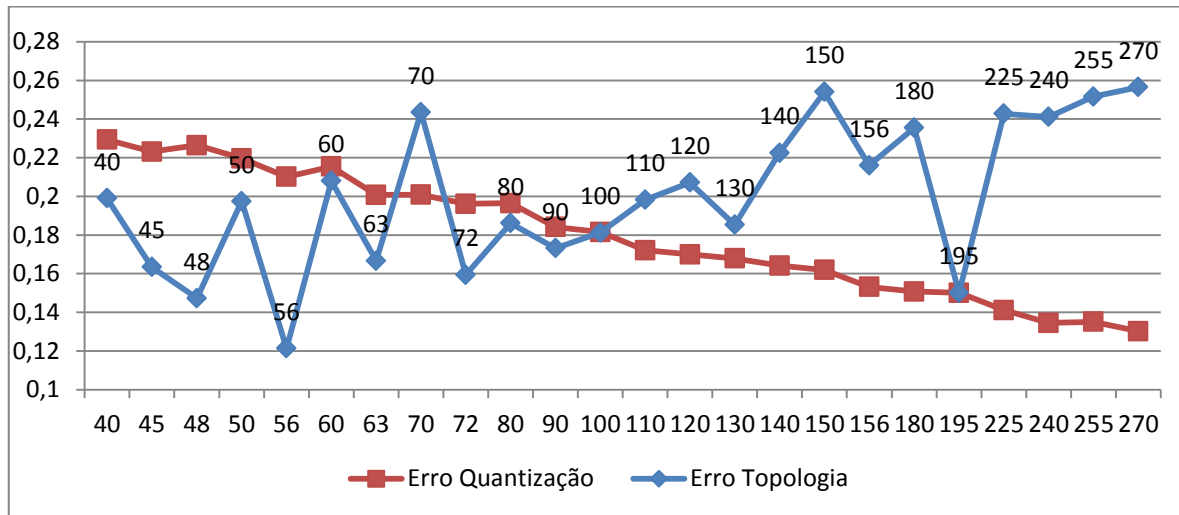


Figura 4.4 - Gráfico de Erros x Número de neurónios – 2002 a 2007

### 4.3. TOPOLOGIA E NORMALIZAÇÃO

A topologia de vizinhança selecionada foi a hexagonal pois permite maior número de vizinhos que a topologia folha e por não se terem verificado diferenças nos resultados dos testes ao se alternar de topologia.

Quanto à forma, a retangular foi aplicada pois é a mais comum para visualização de dados numa projeção bidimensional. E ainda em relação à função de vizinhança, após testes iniciais aplicados ao conjunto de dados deste trabalho, a função gaussiana mostrou melhores resultados e menor erro de topologia.

Em relação à quantidade de épocas, foram utilizados os parâmetros *default* da ferramenta GeoSOM Suite, nomeadamente 10 iterações na fase de treino e 10 na fase de adaptação. Estes valores padrão foram mantidos pois ao aumentar a quantidade de épocas em qualquer das fases houve diminuição pouco significativa do erro de quantização mas houve ligeiro aumento do erro de topologia. Quanto à apresentação dos padrões de entrada foi selecionada a ordem linear do conjunto de dados e processamento batch.

Já quanto à normalização, foram testados todos os tipos disponíveis no GeoSom Suite, tendo a Z-Score e Min-Max apresentado melhores resultados para o conjunto de dados deste trabalho. Entre estes dois tipos, a normalização Min-Max ('Range') apresentou menor erro de quantização e topologia que o Z-Score ('Var').

A normalização Min-Max opera uma transformação linear da amplitude original dos valores de entrada, para uma amplitude pré-determinada (tipicamente entre 0-1). O valor mínimo da distribuição inicial  $\min_1$  é transformado para um novo mínimo  $\min_2$ , o mesmo acontecendo com o valor máximo  $\max_1$  que passa a  $\max_2$ . Todos os pontos entre estes dois valores são transformados linearmente para a nova escala. Uma das vantagens deste procedimento consiste na preservação de todas as relações entre os valores, não introduzindo qualquer enviesamento nos dados, ou seja, a forma do histograma é preservada. Este método de normalização é muito usado e constitui um bom padrão de comparação para outros métodos (Bação, 2011)

Já a normalização Z-Score transforma os valores da variável de entrada de forma que a média seja 0 e a variância 1. O primeiro passo consiste no cálculo da média e desvio-padrão dos dados de entrada. Em seguida, subtrai-se a cada entrada o valor da média e divide-se pelo desvio padrão. A normalização pelo Z-Score funciona particularmente bem quando desconhecemos os mínimos e máximos das variáveis de entrada, ou quando existem *outliers* que inviabilizam a normalização pelo Min-Max (Bação, 2011)

No conjunto estudado, em particular, os períodos de crise não devem ser interpretados como *outliers* pois apresentam características importantes que devem ser consideradas no estudo, por isso o tipo de normalização selecionado foi o Min-Max.

#### **4.4. RESULTADOS SELF-ORGANIZING MAP**

##### **4.4.1. Resultados Parciais**

Antes de apresentar os resultados finais do trabalho, serão discutidos alguns resultados parciais. Muitos foram os testes realizados, principalmente para definição do conjunto de variáveis e tamanho apropriado da matriz (número de neurónios).

Um dos testes parciais importante para determinar os resultados finais foi a consideração ou não da variável preço no conjunto final de variáveis. Nos primeiros testes realizados com os mapas auto-organizáveis, selecionou-se as variáveis a seguir:

- **PX LOW:** Preço mínimo diário
- **PX\_VOLUME:** Número de ações negociadas em 1 dia
- **IX\_HURST\_30:** Índice de Hurst calculado com intervalos de 30 dias

- **VAR\_PRECO\_1:** Variação do preço 1 dia que representa a flutuação diária do preço mínimo
- **VAR\_PRECO\_20:** Variação do preço 20 dias
- **VAR\_PRECO\_30:** Variação do preço 30 dias
- **VAR\_PRECO\_60:** Variação do preço 60 dias
- **VAR\_PRECO\_100:** Variação do preço 100 dias
- **VAR\_PRECO\_120:** Variação do preço 120 dias
- **VAR\_PRECO\_180:** Variação do preço 180 dias

Com base nestas variáveis, foi selecionado um subconjunto de dados referente ao período de 2002 a 2011, composto por 2376 observações. Este período foi selecionado pois inicialmente pretendia-se dividir o conjunto de 10 anos em períodos de estabilidade e crise, nomeadamente 2002 – 2007 como estabilidade, 2007 – 2009 como crise e 2009 – 2011 como recuperação, e desse modo o período de 2002 a 2011 englobava os três subconjuntos pretendidos.

A análise dos erros de topologia e quantização revelou que a matriz 15x12 (180 neurónios) apresentava uma boa representação dos dados, de acordo com os erros de topologia e quantização que se apresenta a seguir, e portanto foi selecionada para os testes parciais.

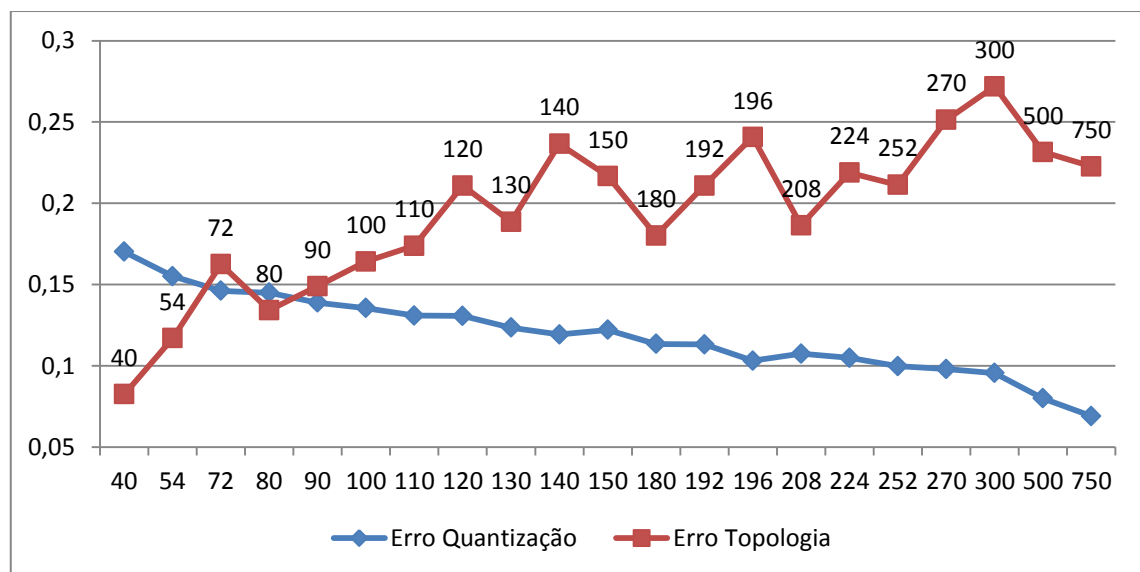


Figura 4.5 - Gráfico de Erros x Número de neurónios – 2002 a 2011

Ao serem analisados os primeiros resultados, verificou-se que os dados apareciam muito ordenados em relação à evolução no tempo. A UMAT (*Unified Distance Matrix*) com *Hits* indica uma grande concentração de dados espalhados pelas bordas da matriz, bem como na zona mais escura do mapa que sugere o surgimento de um *cluster*.

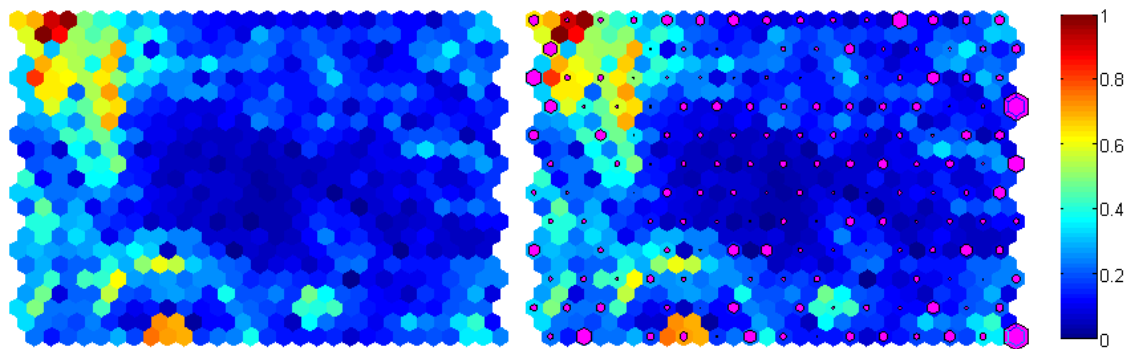


Figura 4.6 - Umat e Umat com *Hits* com Var. Preço: Período 2002 a 2011

A figura a seguir (4.7) representa a distribuição dos dados pela matriz UMAT. Os dados foram agrupados por meses no formato “mmaa”, por exemplo: jan01 significa Janeiro-2001. Para construção desta matriz foi utilizada a função `som_autolabel` com parâmetro 'vote', isto significa que o mês com maior número de casos é escolhido para representar o neurónio.

Nesta Matriz de *Labels* é possível identificar a característica de ordenação dos dados na distribuição da matriz, uma vez que vários meses próximos no tempo também aparecem próximos na matriz.

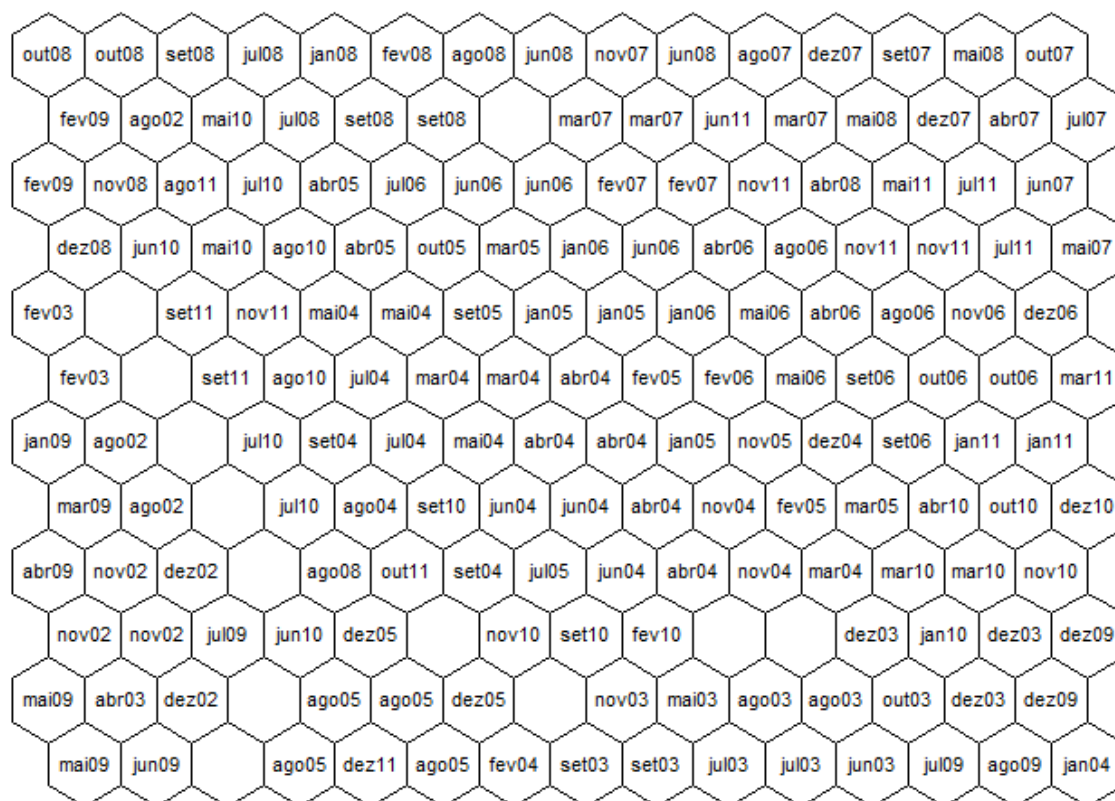


Figura 4.7 – Matriz de *Labels* com Var. Preço: Período 2002 a 2011

Por outro lado, a análise do mapa processado sem a variável preço, revelou uma melhor segmentação dos dados, uma vez que os dados referentes aos períodos de crise aparecem melhor isolados na parte inferior da matriz UMAT, encontrando-se os demais períodos distribuídos pela grande área azul que aparece no mapa.

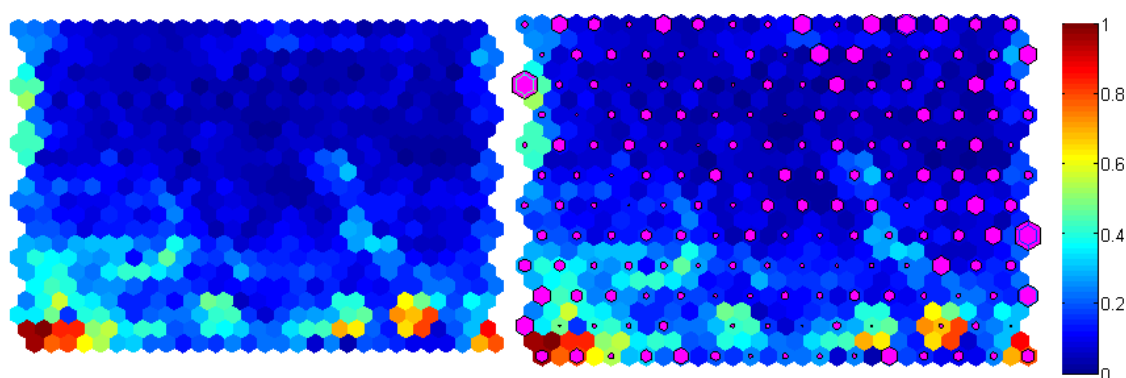


Figura 4.8 - Umat e Umat com *Hits* sem Var. Preço: Período 2002 a 2011

A matriz Matriz de *Labels* apresenta a distribuição dos dados ao longo da matriz, onde se pode verificar que os dados não estão tão ordenados quanto o observado na matriz que incluía a variável preço.

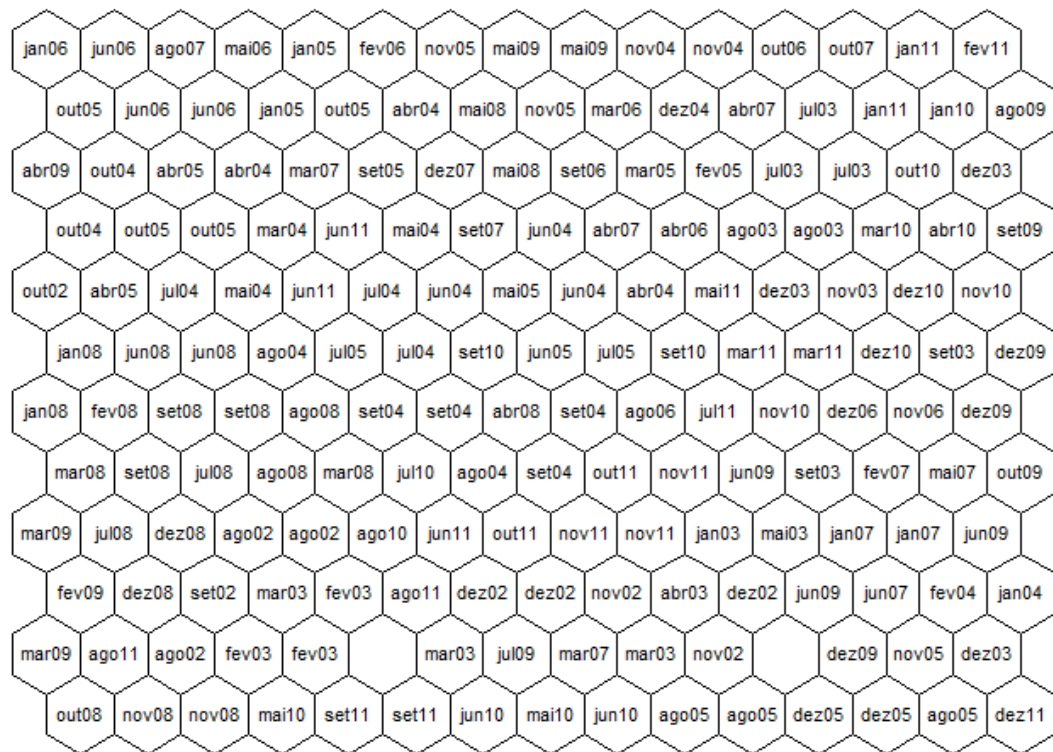


Figura 4.9 – Matriz de *Labels* sem Var. Preço: Período 2002 a 2011

Dessa forma, optou-se por excluir a variável preço do conjunto por esta determinar a ordenação dos dados e prejudicar a segmentação.

Em relação à divisão do período de dados em subconjuntos, após vários testes, concluiu-se que esta divisão não apresentava ganhos nos resultados, pois como os períodos apresentam características similares e os mapas são segmentados com base nas diferenças, é importante determinar diferenças mais significativas para obtenção de melhores resultados, ou seja, englobar períodos de estabilidade e crise em conjunto para os conseguir distinguir.

Entretanto, serão apresentados os resultados respeitantes ao período de estabilidade 2002 a 2007 para comparação com os resultados do conjunto de 10 anos.

#### 4.4.2. Período de 2001 a 2011

A primeira análise foi realizada no conjunto compreendido entre Janeiro-2001 a Dezembro-2011, como foi abordado no capítulo 3.2. Este período passou por duas crises no mercado financeiro, a primeira em 2002 e a segunda compreendida entre 2008 e 2009.

Na Figura 4.10, são apresentadas as matrizes de distâncias UMAT e a UMAT com *Hits*, que representam a distribuição dos dados através da matriz. Pode-se verificar uma grande zona em azul-escuro e algumas zonas mais claras espalhadas pela matriz, além de uma concentração de zonas com escala mais alta no lado esquerdo que cobre principalmente os cantos superiores e inferiores. No canto inferior esquerdo, as regiões em vermelho representam os períodos de crise mais graves, como Setembro-2002 e Novembro-2008. Esta informação pode ser visualizada na Figura 4.11 – Matriz de Labels.

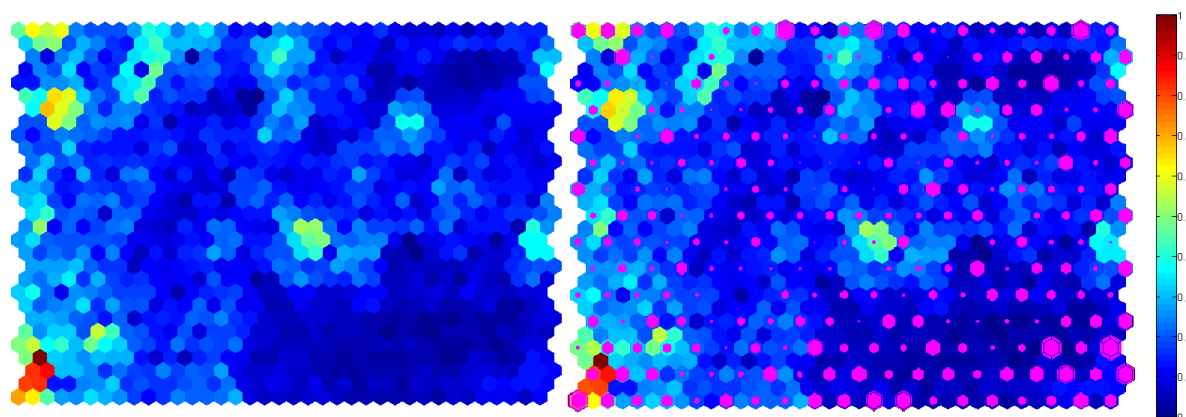


Figura 4.10 - Umat e Umat com Hits: Período 2001 a 2011

A figura a seguir (4.11) apresenta a distribuição dos dados pela Matriz de *Labels*. Nessa matriz é possível visualizar como os meses estão distribuídos ao longo a UMAT.



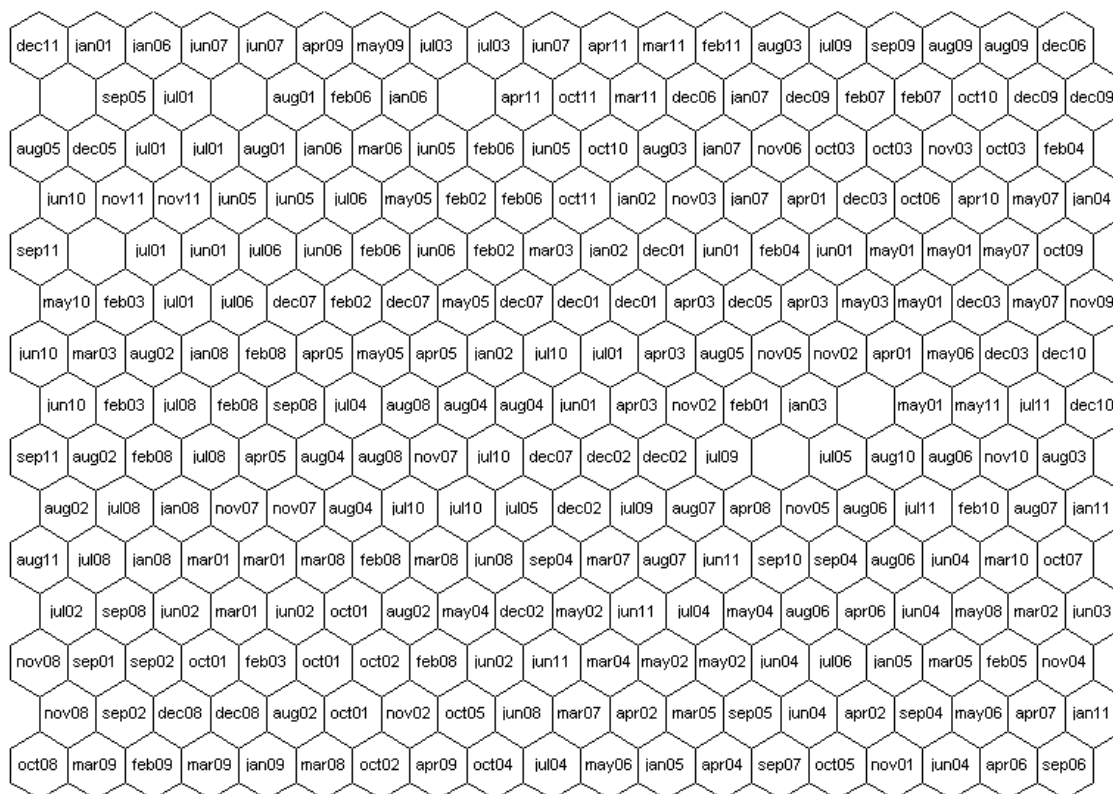


Figura 4.11 – Matriz de Labels: Período 2001 a 2011

Outra forma de compreender a UMAT é através da análise das *component planes*. Estas representam as variáveis utilizadas na construção do modelo. Na Figura 4.12 podem ser observadas as 12 componentes utilizadas neste experimento.

A `PX_VOLUME` representa o volume e `VOLUME_MEDIO_10` representa a variação do volume em relação aos últimos 10 dias. Estas variáveis apresentam excelentes resultados e confirmam as teorias da Análise Técnica de que o volume sustenta o movimento dos preços. Segundo Deschatre & Majer (2006), as regras básicas para os volumes são:

1. Quando preços e volume sobem, os preços tendem a aumentar ainda mais;
2. Quando os preços estão a subir e o volume está a diminuir, a tendência de alta está para ser revertida (geralmente ocorre uma diminuição no ritmo de aumento de preços para então ocorrer a queda);
3. Quando os preços estão a cair e o volume está a subir, a tendência é de diminuição de preços;

4. Quando preços e volume estão a cair, a tendência de baixa está para ser revertida (geralmente ocorre uma diminuição no ritmo da queda dos preços para então haver um aumento).

A VOLUME\_MEDIO\_10 confirma a teoria do item 1 para o conjunto de dados que se localiza na região a vermelho na parte superior do centro-direita e na parte inferior direita. Nestas regiões, estão concentrados os momentos de alta dos preços e momentos em que os preços ainda tinham tendência a aumentar, como Agosto-2003, Setembro-2006, Dezembro-2009, entre outros.

Os dados concentrados no canto inferior esquerdo de ambas as variáveis de volume sustentam o item 3, ou seja, os períodos de maior queda dos preços foram os momentos em que houve maior aumento no volume. Certamente os investidores quiseram aproveitar a queda dos preços para aumentar a carteira de ativos.

Os itens 2 e 4 estão representados nas zonas a azul, principalmente na variável que representa a variação do volume, mas encontram-se em lados opostos. As zonas azuis observadas no canto inferior esquerdo revelam enfraquecimento do movimento de queda e inversão de tendência para alta, o que pode ser exemplificado pelo período Dezembro-2008, revelando aqueles que surgem do lado direito do mapa uma inversão de tendência de alta para queda, como é o caso de Setembro e Outubro de 2007.

As variáveis de variação do preço (VAR\_PRECO, figura 4.12) sugerem a distribuição dos preços ao longo da matriz UMAT. As zonas que aparecem a azul representam que o preço atual encontra-se muito abaixo da média dos últimos  $n$  dias<sup>4</sup>, e as zonas que aparecem a vermelho indicam que o preço atual está bastante superior à média. Pode-se então constatar que os preços mais baixos estão concentrados no lado esquerdo do mapa. A diferença entre as variáveis de curto, médio e longo prazo pode então ser verificada pela maior concentração da zona vermelha no canto superior direito do mapa à medida que se aumenta o período da média móvel. Isto ocorre pelo fato de ao ser considerado um período maior para a média, as variações tendam a ficar menores. Na variação 180 dias (VAR\_PRECO\_180) observa-se uma variação suave nas cores do mapa, aparecendo no canto superior direito a maior concentração das variações de preço em relação à média móvel 180 dias e no canto inferior esquerdo estão concentradas as menores variações, ou seja, os momentos em que o valor atual do índice surgia muito mais abaixo da média dos últimos 180 dias.

---

<sup>4</sup> Onde  $n = \{1, 10, 20, 30, 60, 100, 120, 180\}$

A variação do preço indica as zonas em que os preços estão subvalorizados (zona azul), com preços médios (zonas verde e amarela) ou supervalorizados (zonas vermelhas). Esta variável por si só não revela grandes detalhes sobre as tendências dos movimentos. No entanto, se for analisada em conjunto com o volume/volume médio, consegue-se perceber que na região do canto superior direito, por exemplo, os preços estão bastante altos, revelando um enfraquecimento da tendência pois houve uma diminuição do volume.

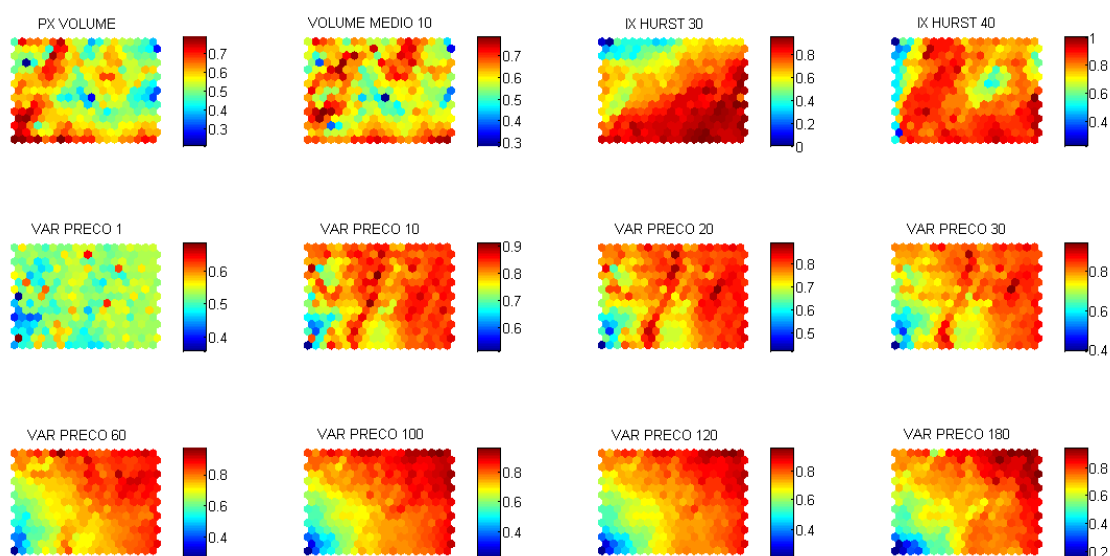


Figura 4.12 - Component Planes: Período 2001 a 2011

As variáveis IX\_HURST\_30 e IX\_HURST\_40 representam a memória da série, isto é, a tendência, para a probabilidade de repetição de um valor. Nas regiões a vermelho são as regiões que apresentam maior probabilidade de a tendência do movimento manter-se. Entretanto ao analisar o comportamento do índice de Hurst para o *Dow Jones Industrial Average* isoladamente (Figura 4.13), pode-se observar que ao longo do período analisado os valores de Hurst situam-se num nível superior a 0,5 indicando que a série apresenta memória persistente. Apenas alguns pontos em Jan-2001, Jul-2003, Mar-2008 e Set-2011 possuem uma oscilação para valores iguais ou inferiores a 0,5, os quais revelam a ausência de memória e existência de passeio aleatório.

Se por um lado este indicador confirma a teoria da Análise Técnica de que o histórico de preços apresenta memória, o que quer dizer que pode ser utilizado para prever o futuro. Por outro lado, este não se apresentou muito significativo na descrição dos movimentos do mercado pois ao comparar as oscilações do índice de

Hurst com o movimento dos preços não parece haver nenhuma relação que contribua para explicar a variação dos preços.

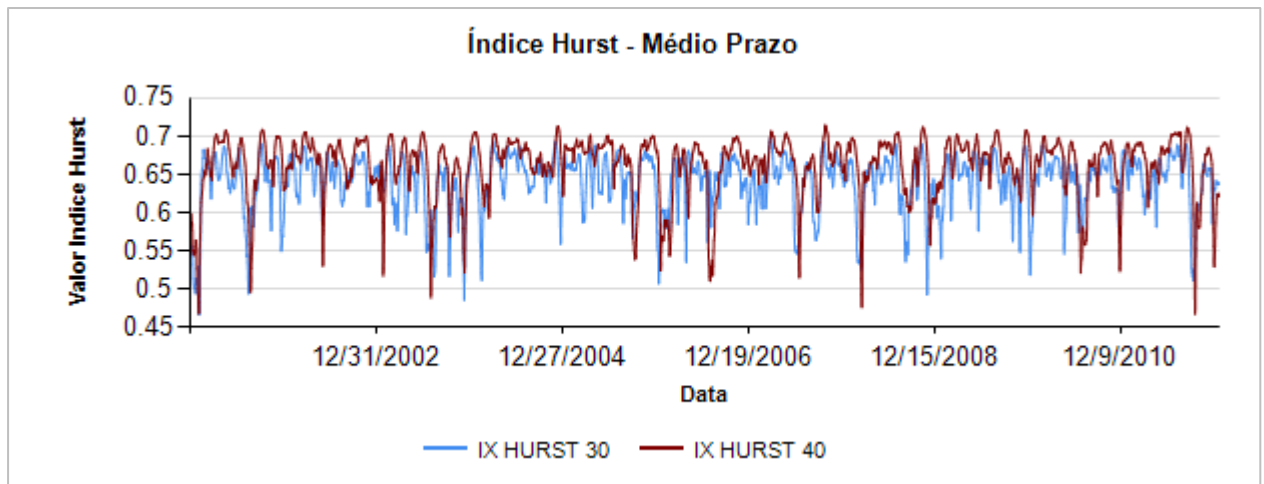


Figura 4.13 - Índice de Hurst Médio Prazo: Período 2001 a 2011

A seguir pode ser visualizada uma representação da segmentação de *clusters* no gráfico dos preços (4.14 e 4.15) para permitir visualizar os resultados da separação realizada pelo modelo. São então apresentadas duas segmentações distintas, uma segmentação dos vales (zonas mais escuras no mapa) e uma outra dos picos (zonas mais claras, que são consideradas como zonas de separação dos vales).

Na Figura 4.14 a UMAT foi dividida em 3 segmentos: Clus 1 representa o segmento nº 1 destacado a amarelo, Clus 2 representa o segmento nº 2 em cor laranja-escuro e o Clus 3 representa o segmento nº 3 em cor vermelho-escuro. O Clus 0 destacado a azul-claro representa as demais zonas do mapa que não foram segmentadas, ou seja, o fundo do mapa.

Como se pode verificar pela análise do gráfico dos preços, a zona da UMAT que aparece a amarelo (Clus 1), está geralmente relacionada a movimentos de alta dos preços. Já as regiões a laranja-escuro (Clus 2), sugerem momentos de estabilidade, onde os movimentos apresentam variações de pequena amplitude. Por outro lado, a zona a vermelho-escuro (Clus 3) aponta para momentos de transição, onde há um enfraquecimento do movimento e uma certa incerteza quanto à continuidade do sentido do movimento, pois aparece em zonas onde o seguimento dos preços se inverte tanto para altas quanto para baixas.

Na Figura 4.15, a UMAT foi dividida em 6 segmentos, sendo que o Clus 0 representa o fundo da matriz e os outros Clus 1 a 6 representam os segmentos numerados de acordo com o mapa. Nesta representação de análise dos picos, nota-se a boa capacidade do modelo de separar os momentos de crise, representados pela cor vermelha (Clus 6). Outra característica interessante é a de que os picos centrais da matriz (Clus 1 e Clus 5) marcam zonas de transição, confirmando a zona de instabilidade que referimos anteriormente na Figura 4.14 (Clus 3). Estes dois segmentos aparecem antes dos movimentos mudarem de sentido, provavelmente marca momentos em que a trajetória muda do lado direito para o lado esquerdo da matriz ou vice-versa.

No gráfico dos preços (4.15) pode-se perceber melhor a ordenação dos dados dentro do mapa. Onde o *cluster* 6 representa claramente os movimentos de queda, embora esta representação de segmento provavelmente não ofereça a melhor separação, uma vez que a ferramenta utilizada não separa automaticamente os segmentos, apenas sugere uma segmentação (conforme se apresentou no capítulo 3.6. Processamento e Análise de Dados). Os demais segmentos de 1 a 5 estão posicionados em momentos estratégicos, geralmente marcando inversão de movimentos.

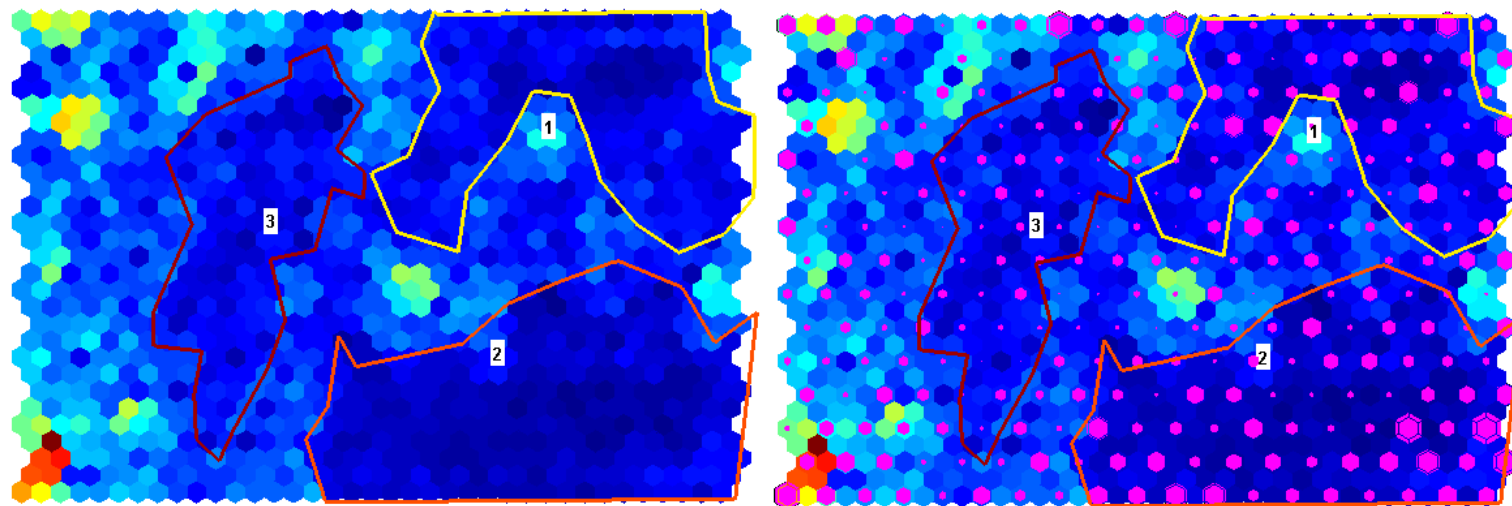
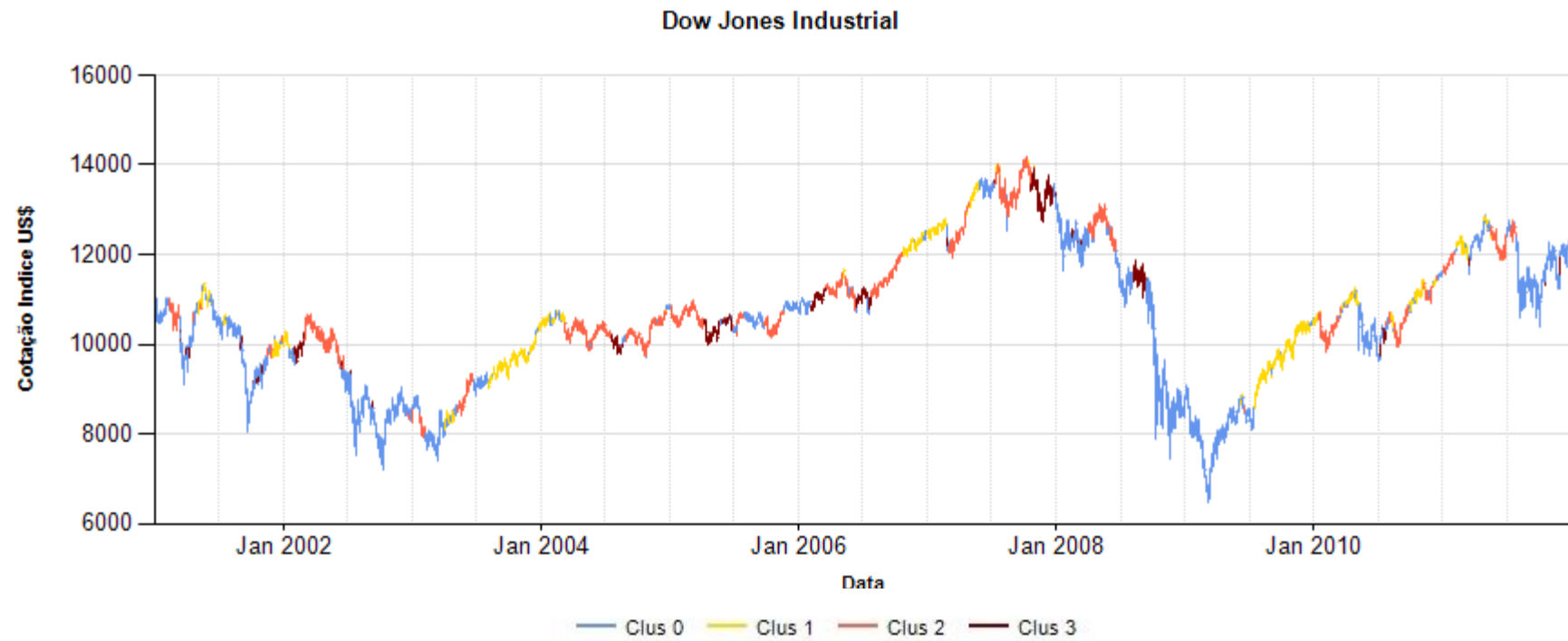


Figura 4.14 - Análise de Cluster Vales: Período 2001 a 2011

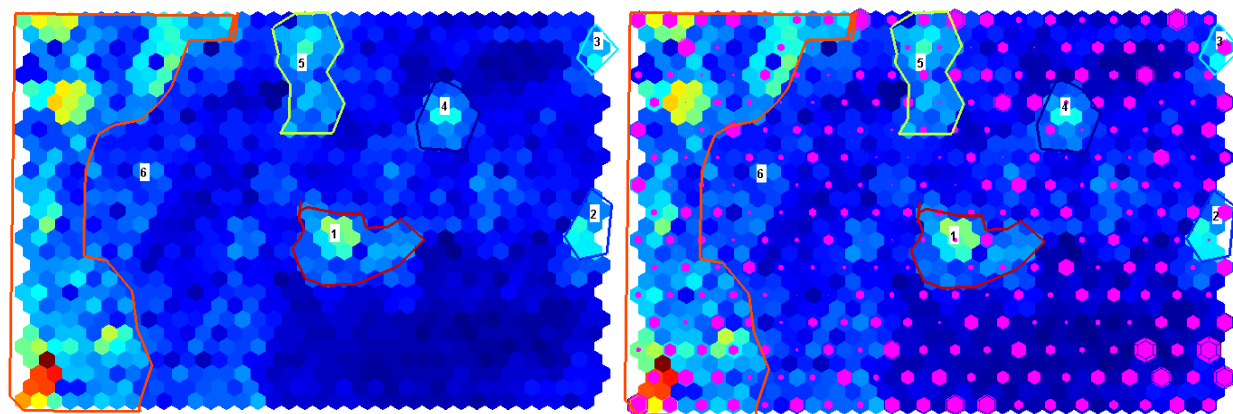
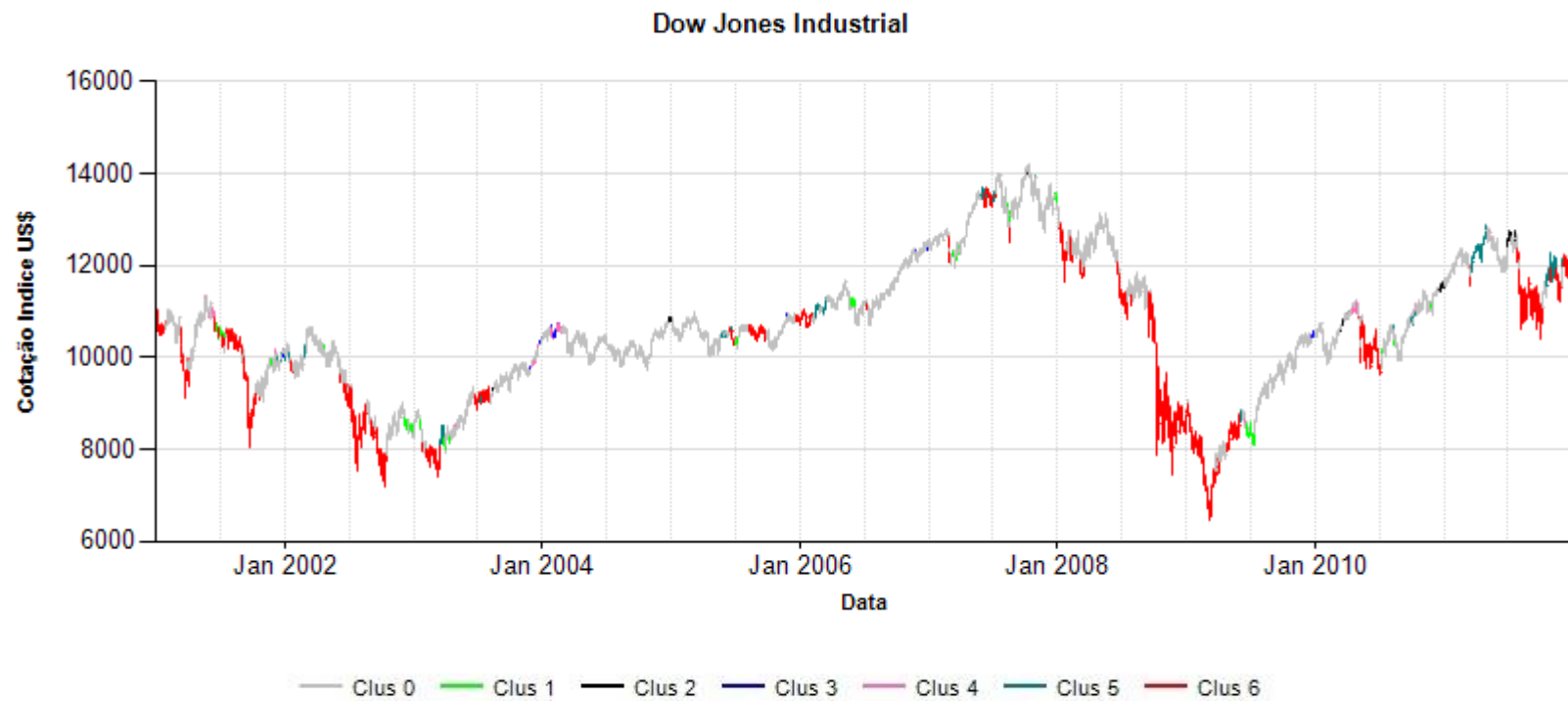


Figura 4.15 - Análise de Cluster Topos: Período 2001 a 2011

#### 4.4.3. Período de 2002 a 2007

Esta análise foi realizada sobre um subconjunto de dados compreendido entre Julho-2002 a Junho-2007, a qual é caracterizada por um momento de estabilidade dos mercados, como se discutiu no capítulo 3.5. Análise de Subconjuntos.

Na matriz UMAT (4.16) deste subconjunto nota-se muito mais variação nas cores do mapa. Como este subconjunto regista um período de certa calma nos mercados e pequena variação nos movimentos dos preços, o mapa apresenta sensibilidade às pequenas variações. Na matriz UMAT com *Hits* percebe-se que os dados ficam concentrados em algumas zonas do mapa, mais genericamente nos cantos da matriz.

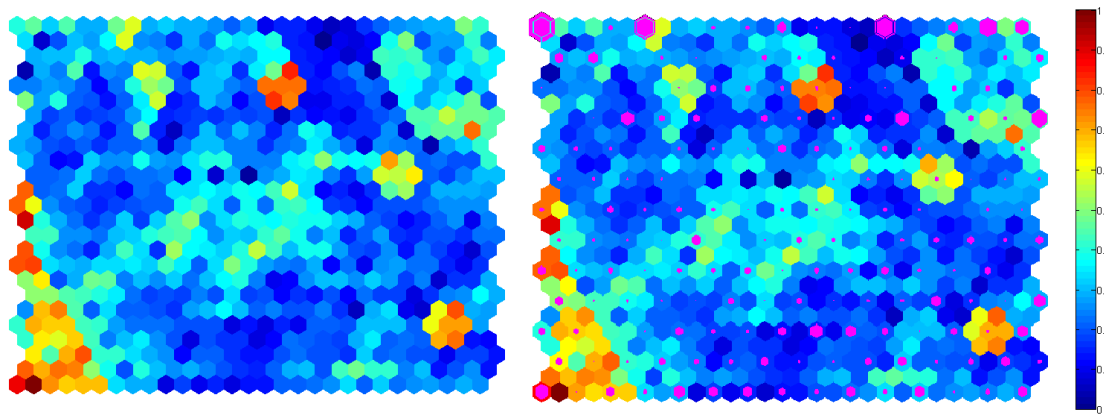


Figura 4.16 - UMAT e UMAT com Hits: Período 2002 a 2007

Mais uma vez, os dados correspondentes aos períodos com preços mais baixos ficaram concentrados no canto inferior esquerdo, como se pode verificar na matriz de *labels*, (Figura 4.17).



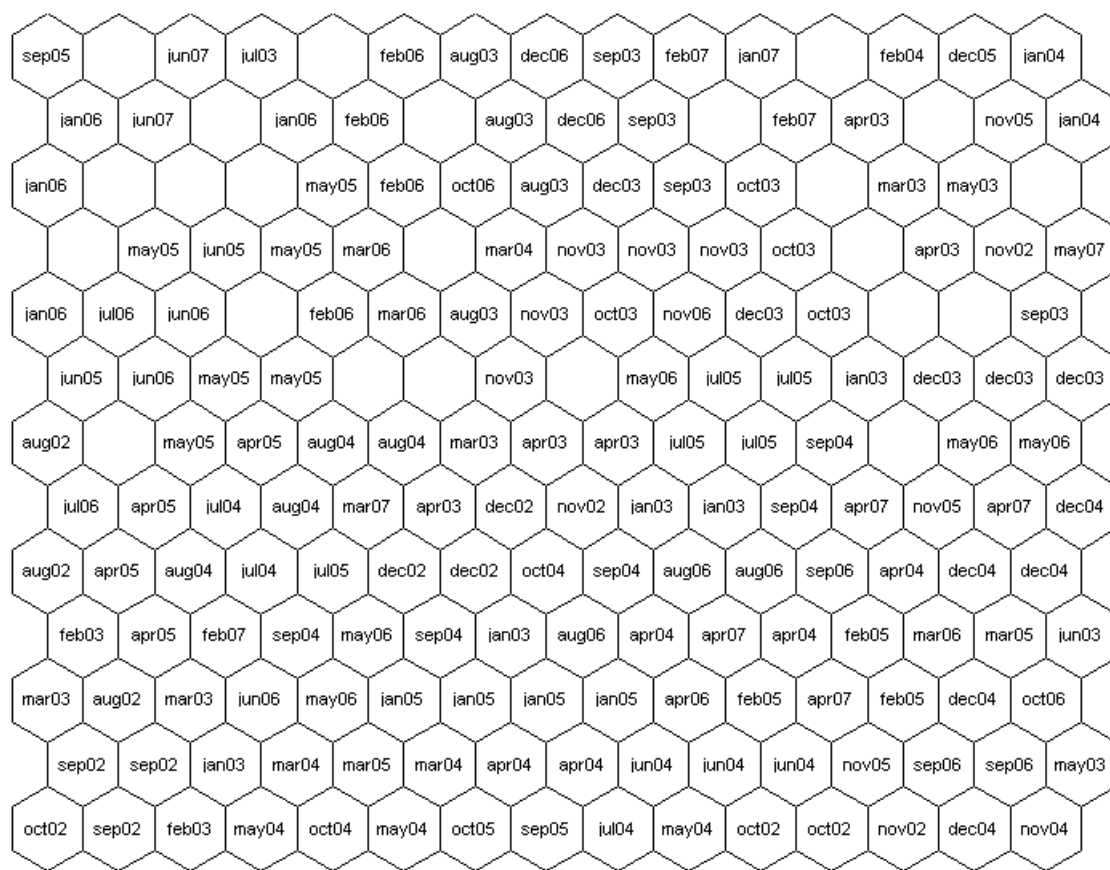


Figura 4.17 – Matriz de Labels: Período 2002 a 2007

Vamos agora analisar as *component planes* para perceber a distribuição dos dados pela UMAT. Pode-se observar que as variáveis seguem as mesmas características apresentadas para o período de 10 anos, com ligeiras diferenças nas intensidades acompanhando o novo conjunto de dados.

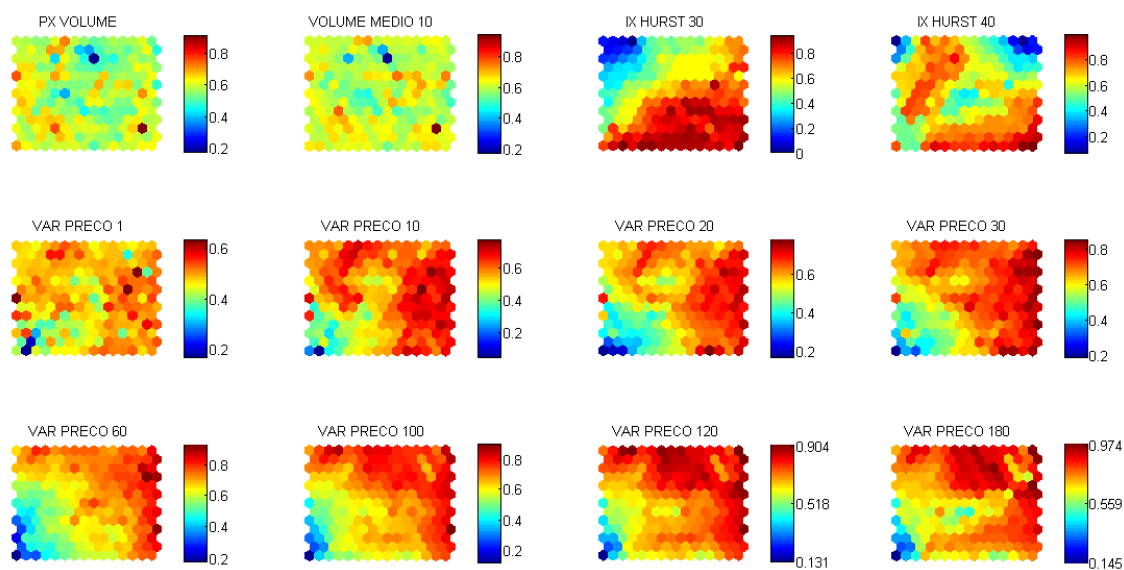


Figura 4.18 - Component Planes: Período 2002 a 2007

As variáveis de volume apresentam poucas variações neste mapa, ainda assim, nota-se que os maiores níveis de volume estão nas regiões do mapa onde há tendências de baixa ou de alta, tomando como exemplo Janeiro-2004 e Novembro-2005 que representam tendências de alta e Fevereiro-2003 que representa tendência de baixa. Já as regiões de baixo volume indicam enfraquecimento do movimento e possibilidade de inversão, como é o caso de finais de Dezembro-2004.

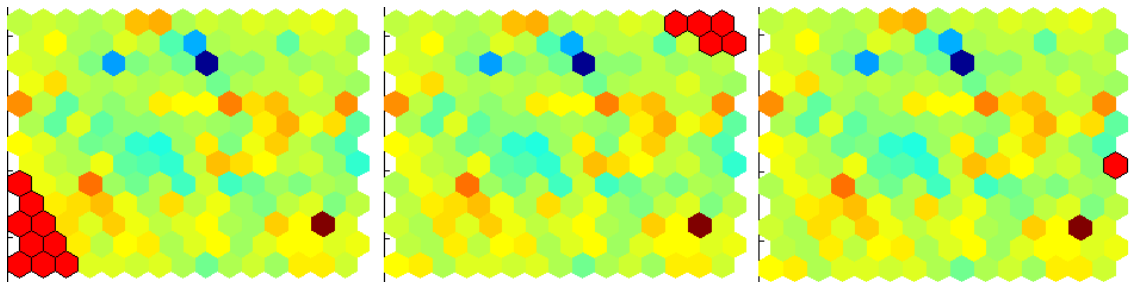


Figura 4.19 - Volume Médio: Tendências de baixa, alta e inversão do movimento

Com o objetivo de visualizar a distribuição do mapa no gráfico dos preços, dividiu-se a UMAT novamente em segmentos. A Figura 4.20 que se segue apresenta a segmentação dos vales.

Ao analisar a segmentação neste mapa, é possível reparar que o modelo separa basicamente a matriz em duas zonas distintas, a parte superior da matriz, onde ficam sobretudo concentrados os momentos de maior estabilidade: Clus 1, 3 e 4 da Figura 4.20 e Clus 2, 4 e 5 da Figura 4.21. Na parte inferior da UMAT, a qual sugere momentos mais instáveis como se pode verificar nos segmentos 2 e 5 da Figura 4.21. Apesar desta apresentar movimentos de alta dentro destes segmentos, estes caracterizam-se por momentos em que o sentido do movimento se altera mais rapidamente do que nos outros segmentos.

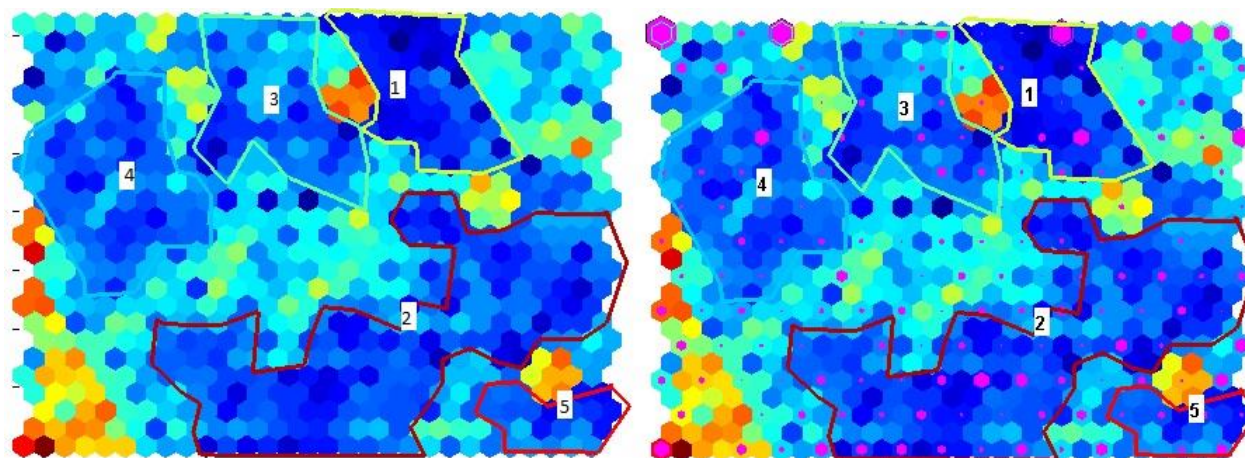
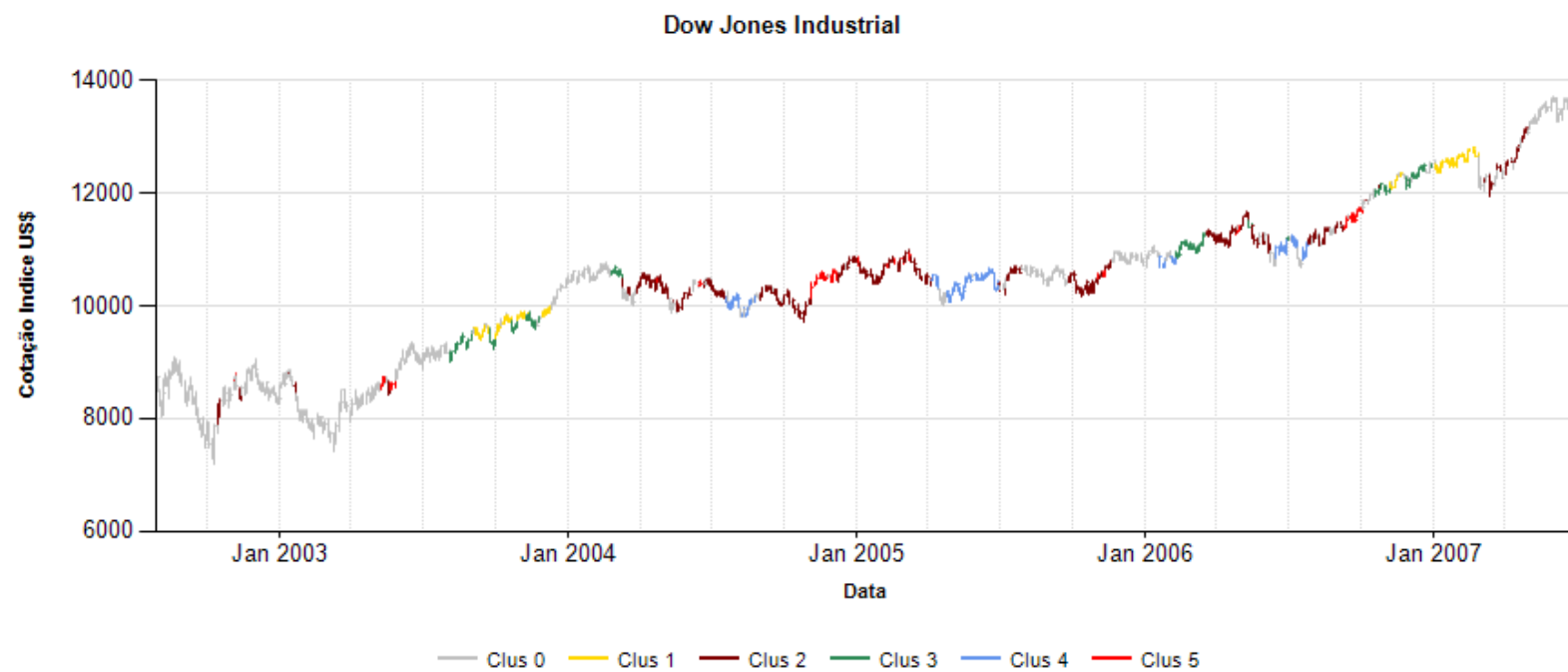


Figura 4.20 - Segmentação Mapa 2002 a 2007

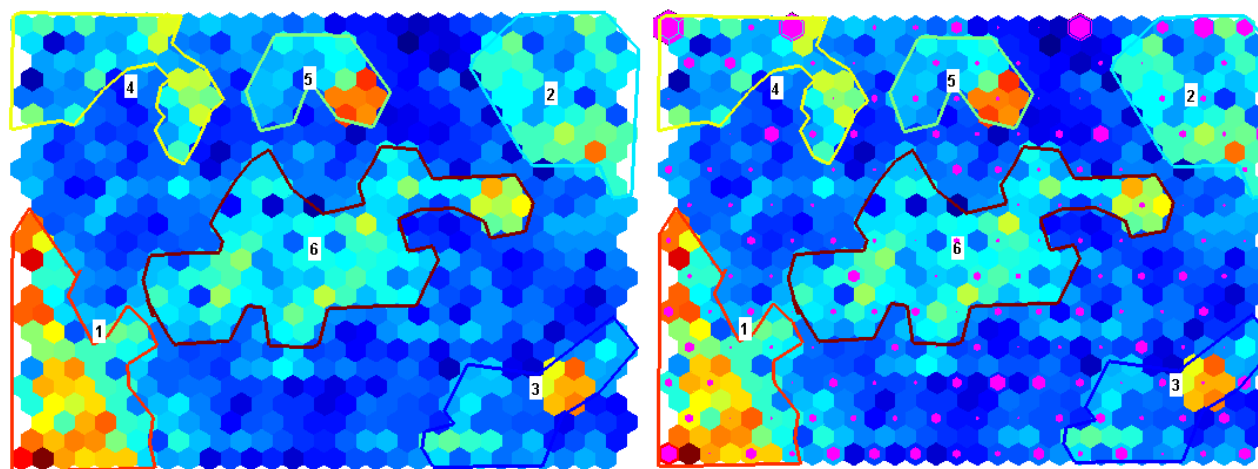
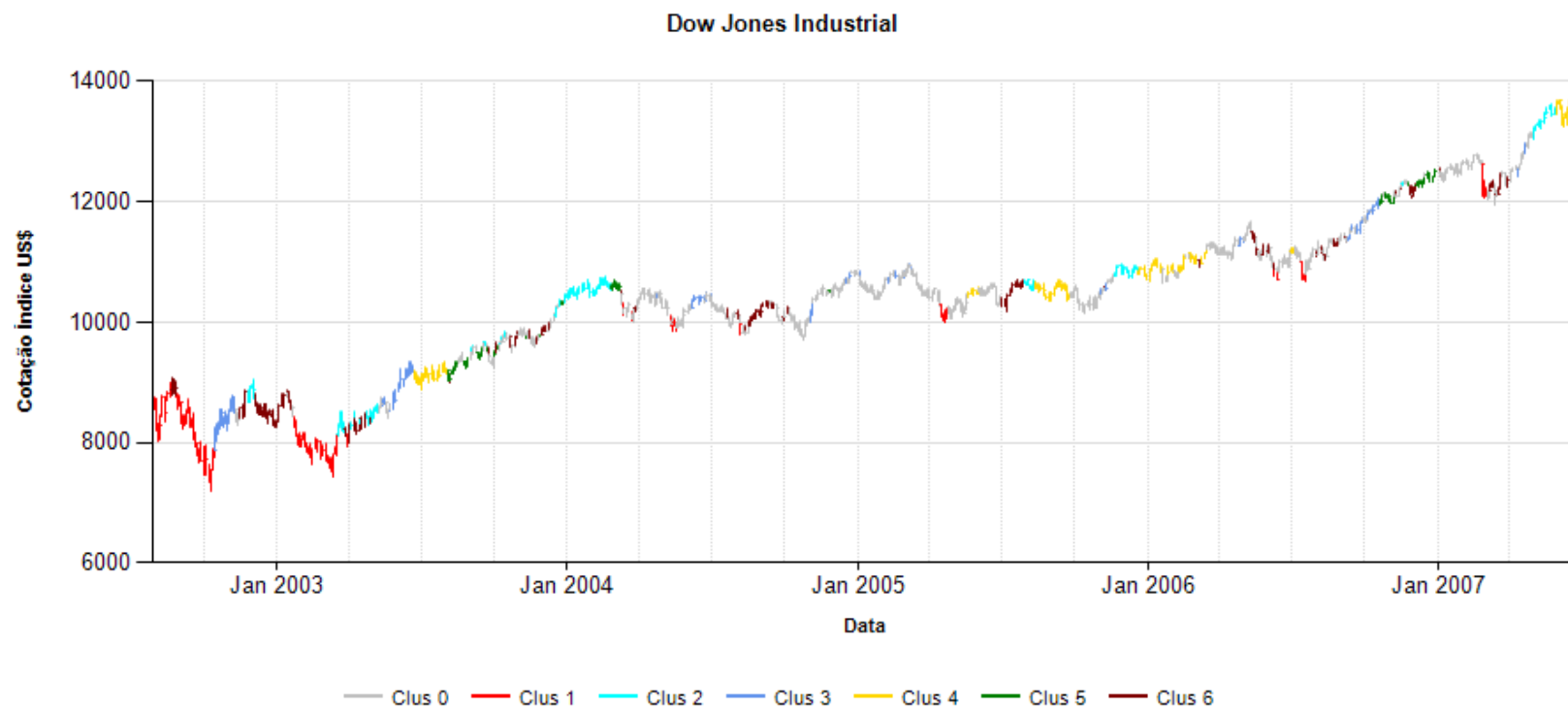


Figura 4.21 - Segmentação Mapa 2002 a 2007

#### 4.5. VALIDAÇÃO DOS RESULTADOS

Para validar os resultados, foi simulado um investimento na bolsa utilizando duas estratégias: uma mais conservadora e outra com maior risco. Os valores apresentados são apenas demonstrativos, pois o índice *Dow Jones Industrial Average* é composto por um conjunto de 30 ações, não é propriamente uma ação comercializada. Seria necessário comprar ação de cada uma das empresas para obter uma carteira de ações representada pelo índice.

A segmentação utilizada para validar os resultados foi a apresentada na Figura 4.14 - Análise de Cluster Vales: Período 2001 a 2011. Na estratégia designada Conservadora, optou-se por comprar e manter o índice nos períodos em que fosse classificado como *Cluster 1* (Amarelo), que normalmente está relacionado com movimentos de alta dos preços. Portanto, nesta estratégia é efetuada a compra ou manutenção do índice enquanto este pertencer no *cluster 1* e a venda ocorre quando o movimento pertencer a outro *cluster*. O resultado utilizando esta estratégia foi um ganho de 51,10% no período de 10 anos.

Já na Estratégia de Risco, optou-se por investir nos *Clusters 1, 2 e 3*, ou seja, investe-se nos momentos de alta dos preços representado pelo *cluster 1*, nos períodos de estabilidade (*cluster 2*) e ainda nos momentos em que o movimento está a perder força e na iminência de ser invertido (*cluster 3*). Em suma, nesta estratégia opta-se por comprar ou manter o investimento sempre que o movimento é classificado num vale e vender sempre que o movimento é classificado no *cluster 0*. Nesta estratégia, embora ao longo do percurso se tenham verificado perdas, o resultado superou a Estratégia Conservadora e o ganho foi de 75,86% no período de 10 anos.

O preço utilizado para os cálculos foi o preço de fechamento. A tabela a seguir apresenta um resumo dos resultados das duas estratégias.

<b>Estratégia Conservadora</b>	Comprar/Manter: Cluster 1 Venda: Clusters 0, 2 e 3
<b>Valor Investido(\$)</b>	10796,65
<b>Lucro (\$)</b>	5517
<b>Lucro (%)</b>	51,1
<b>Estratégia de Risco</b>	Comprar/Manter: Clusters 1, 2 e 3 Venda: Cluster 0
<b>Valor Investido(\$)</b>	10781,45
<b>Lucro (\$)</b>	8178,31
<b>Lucro (%)</b>	75,86

Tabela 4.1 - Tabela Estratégias de Investimento

A figura a seguir (4.22) mostra a evolução do capital líquido (lucro) no fechamento de cada mês. Nos meses em que não houve transações de compra e venda, manteve-se o valor de capital do mês anterior.

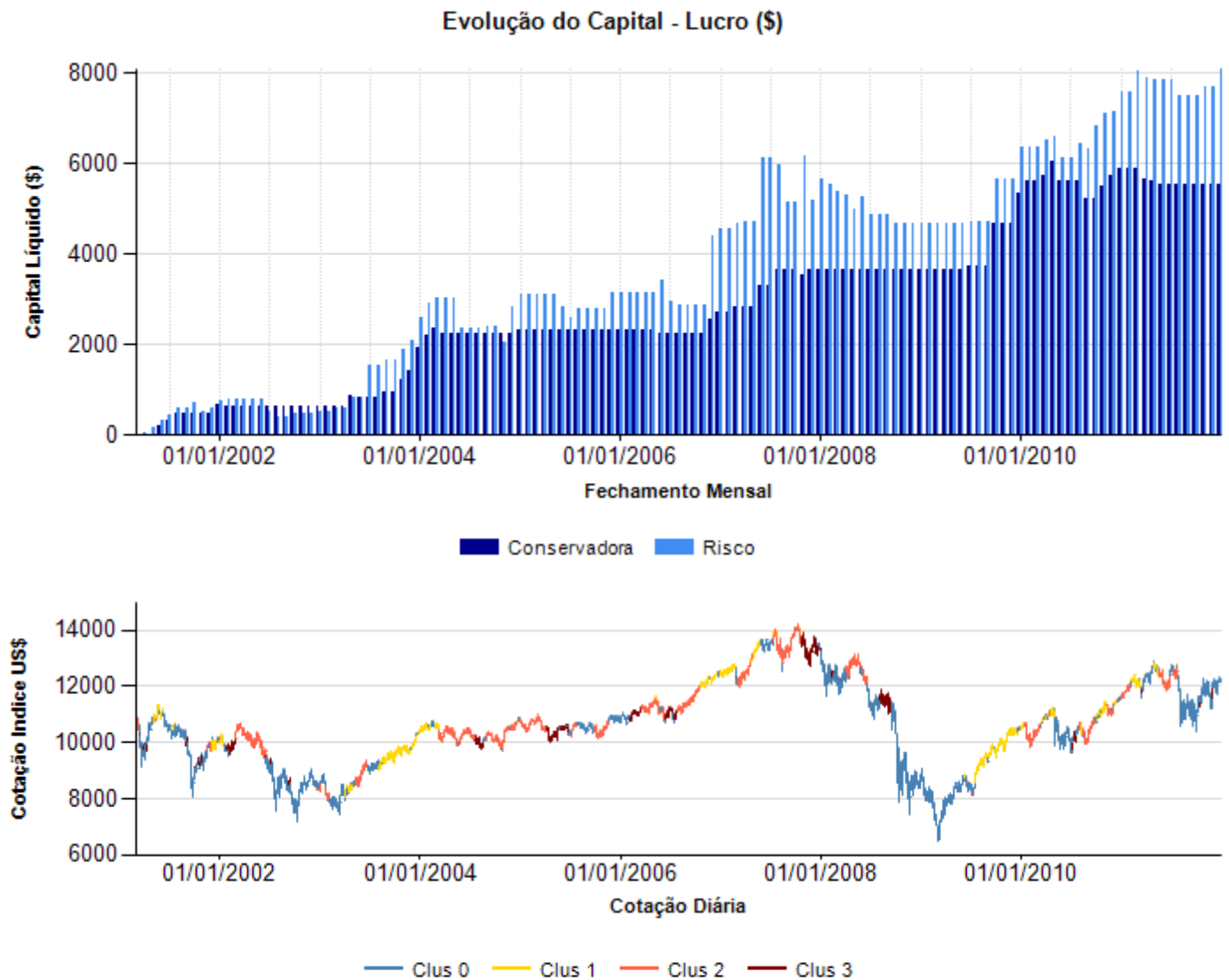


Figura 4.22 – Evolução do capital líquido no fechamento mensal

Embora os resultados apresentados não sejam certamente os melhores em relação aos lucros de investimento, estes mostraram-se bastante satisfatórios ao considerar que são resultados de uma rede não supervisionada e também foram importantes para confirmar que os mapas auto-organizáveis também podem ser

aplicados no segmento do mercado financeiro evidenciando grande poder de descoberta de padrões, o que neste trabalho contribuiu para sinalizar os momentos de estabilidade e crise nos mercados.

Com qualquer outra estratégia que mantenha os investidores fora do mercado em momentos de muita instabilidade ou crise, verificou-se com a utilização destas duas estratégias: conservadora e de risco, que é possível obter ganhos no mercado de ações.

## 5. CONCLUSÕES, LIMITAÇÕES E RECOMENDAÇÕES

As componentes mais significativas para compreender os resultados da ordenação do mapa foram o volume, volume médio e a variação do preço.

O volume/volume médio em conjunto com o preço, como foi discutido no capítulo anterior, fornece indícios sobre a intensidade dos movimentos, podendo o volume ser utilizado para confirmar a força de uma tendência. Ao analisar o plano de componentes utilizadas no modelo dos mapas auto-organizáveis, verificou-se a sustentação das teorias da Análise Técnica em relação ao papel do volume na descrição do comportamento do mercado financeiro.

A variação do preço também se apresentou bastante significativa na análise das tendências quando utilizada em conjunto com o volume ou volume médio, pois esta variável permite revelar só por si os momentos em que preços estão subvalorizados, com preços médios ou supervalorizados em relação à média dos últimos  $n$  dias, onde  $n$  pode ser um período de curto, médio ou longo prazo, dependendo do interesse do estudo. Quando empregada em conjunto com o volume/volume médio, consegue-se perceber os momentos em que ocorre o enfraquecimento das tendências.

Já o índice de Hurst para o índice *Dow Jones Industrial Average*, não se mostrou muito significativo na descrição do comportamento do mercado, uma vez que as oscilações do índice não indicaram nenhuma relação com a variação dos preços. Entretanto, os resultados deste indicador apontaram para a confirmação de memória na série, pois os valores mantiveram-se preponderantemente superiores a 0,5 o que indica que a série é persistente.

Quanto aos intervalos mais significativos para cálculo das variáveis, embora neste trabalho se tenha utilizado variáveis de curto, médio e longo prazo, acredita-se que para as variáveis relacionadas com preço os períodos médio e longo prazo são mais indicados, uma vez que nestes intervalos obtém-se maior visibilidade do passado para comparação com o presente.

Em relação às correlações entre as variáveis, conforme se discutiu no capítulo 4.1. Redução das Variáveis, as variáveis preço estão altamente relacionadas com as médias, recomendando-se portanto trabalhar com as médias pois estas contêm mais informações do que o preço diário já que estão relacionadas com um determinado período de tempo.



Por fim, na Validação dos Resultados verificou-se que os mapas auto-organizáveis são um método bastante eficiente para segmentação dos dados relacionados com o histórico de preços do índice DJI, permitindo inclusive distinguir os momentos de estabilidade e os momentos de crise nos mercados, o que na prática seria bastante útil aos investidores pois permitir-lhes-ia criar estratégias de proteção para os momentos de crise.

Contudo, apesar deste trabalho ter contribuído para mapear variáveis que são significativas na descrição do comportamento dos mercados, não foi possível testar um grande conjunto de variáveis pela limitação do tempo. Tendo-se focado apenas nas variáveis mais utilizadas na Análise Técnica, e no teste do índice de Hurst, o qual é importante para determinar a memória de uma série.

Porém existem ainda outras variáveis que são reconhecidas pela Análise Técnica, as quais poderiam ser incluídas em novos estudos: média móvel exponencial (EMA), a média móvel de convergência e divergência (MACD), linha estocástica (KD), índice de força relativa (IFR), que não foram incluídas neste trabalho.

## 6. BIBLIOGRAFIA

- Alhoniemi E., Himberg J., Parhankangas J. & Vesanto J. (08 de 2012). *Laboratory of Computer and Information Science*. Obtido de <http://www.cis.hut.fi/projects/somtoolbox/>
- Baço F. L. (2011). Módulo V – Self-Organizing Maps. *Unpublished manuscript*. Lisboa, Portugal: Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa.
- Basu S. (1977). Investment Performance of Common Stocks in Relation to Their Price-Earnings Ratios: A Test of the Efficient Market Hypothesis. *The Journal of Finance*, pp. 663-682.
- Blume L., Easley D. & O'Hara, M. (1994). Market Statistics and Technical Analysis: The Role of Volume. *The Journal of Finance*, pp. 153-181.
- Brown S. J., Goetzmann W. N. & Kumar, A. (1998). The Dow Theory: William Peter Hamilton's Track Record Reconsidered. *The Journal of Finance*, Vol. 53, No. 4, *Papers and Proceedings of the Fifty-Eighth Annual Meeting of the American Finance Association*, pp. 1311-1333.
- Cerbasi G. (2008). *Investimentos inteligentes*. Rio de Janeiro: Thomas Nelson Brasil.
- Deboeck G. J. & Kohonen, T. (1998). *Visual Explorations in Finance: with Self-Organizing Maps*. New York: Springer.
- Deschatre G. (2009). *INVESTIMENTO EM AÇÕES: PARA OS MOMENTOS DE CRISE E DE CRESCIMENTO*. Rio de Janeiro: Thomas Nelson Brasil.
- Deschatre G. A. & Majer, A. (2006). *Aprenda a Investir com Sucesso em Ações - Análise Técnica e Fundamentalista*. Ciencia Moderna.
- Dow C., Russell R., Carlson C., Shread P. & Sether L. (2009). *Dow Theory Unplugged: Charles Dow's Original Editorials and Their Relevance Today*. W&A Publishing.
- Esa Alhoniemi Johan Himberg, J. P. (s.d.). SOM Toolbox for Matlab 5.
- Fama E. F. (1965). The Behavior of Stock-Market Prices. *The Journal of Business*, pp. 34-105.

- Fama E. F. (1970). Efficient Capital Markets: A Review of Theory and Empirical Work. *The Journal of Finance*, pp. 383-417.
- Ferreira G. (04 de Dec de 2011). *Livro da Bolsa Vol. 1*. Obtido de <http://www.livrodabolsa.com/>
- Greenspan A., Mankiw N. G. & Stein J. C. (2010). The Crisis [with Comments and Discussion]. *Brookings Papers on Economic Activity*, 201-261.
- Hurst H. E. (1951). LONG-TERM STORAGE CAPACITY OF RESERVOIRS. *Transactions of the American Society of Civil Engineers*, pp. 770-799.
- ISEGI-UNL. (2013). GeoSOM. Obtido em 17 de Abril de 2013, de <http://www.isegi.unl.pt/labnt/geosom/index.htm>
- Kaski S., Venna J. & Kohonen T. (2000). Coloring that Reveals Cluster Structures in Multivariate Data. *Australian Journal of Intelligent Information Processing Systems*, 6: 82–88.
- Khan A. U., Bandopadhyaya T. K. & Sharma S. (2008). Classification and Identification of Stocks using SOM and Genetic Algorithm based Backpropagation Neural Network . *IIT: 2008 INTERNATIONAL CONFERENCE ON INNOVATIONS IN INFORMATION TECHNOLOGY*, 81-85.
- Kiviluoto K., & Bergius P. (1997). Analyzing Financial Statements with the Self-Organizing Map. in *Proc. WSOM'97, Workshop Self-Organizing Maps*, (pp. 362-367).
- Kohonen T. (1982). Self-Organized Formation of Topologically Correct Feature Maps. *Biological Cybernetics*, pp. 59-69.
- Kohonen T. (2001). *Self-Organizing Maps*. Springer.
- Li S.T. & Kuo S.C. (2008). Knowledge discovery in financial investment for forecasting and trading strategy through wavelet-based SOM networks. *EXPERT SYSTEMS WITH APPLICATIONS*, 935-951.
- Maimon O. & Rokach L. (2005). *Data Mining and Knowledge Discovery Handbook*. Springer.

- Mandelbrot B. B. (1971). When Can Price be Arbitraged Efficiently? A Limit to the Validity of the Random Walk and Martingale Models. *The Review of Economics and Statistics*, 53(3), 225-236.
- Mandelbrot B. & Van Nessi J. (1968). Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10(4), 422–437.
- Mandelbrot B. & Wallis J. (1969). Robustness of the rescaled range R/S in the measurement of noncyclic long run statistical dependence. *Water Resources Research*, 5(5), 967–988.
- Marcus A., Bodie Z. & Kane A. (2007). *Fundamentos de investimentos*. Bookman.
- Meyers T. A. (2002). *The technical analysis course: a winning program for investors & traders* (Vol. 3rd. Ed). McGraw-Hill.
- Resta M. (1998). An hybrid neural network system for trading financial markets. Em G. J. Deboeck, & T. Kohonen, *Visual Data Explorations in finance with Self Organizing Maps* (pp. 106-116). Springer Verlag London.
- Rhea R. (1932). *The Dow Theory*. New York: Barron.
- Silva B., Panosso G. & Marques N. C. (2012). Applying Neural Networks for Concept Drift Detection. *In conjunction with the 20th European Conference on Artificial Intelligence (ECAI 2012)*.
- Su S. & Huang S. (2003). Applications of model-free estimators to the stock market with the use of technical indicators and non-deterministic features. *JOURNAL OF THE CHINESE INSTITUTE OF ENGINEERS*, 21-36.
- Thomsett M. C. (2006). *Getting Started in Fundamental Analysis*. John Wiley & Sons.
- Thomsett M. C. (2007). *The Stock Market Investor's Pocket Calculator: A Quick Guide to All the Formulas and Ratios You Need to Invest Like a Pro*. Chicago: AMACOM.
- Ultsch A. & Moerchen, F. (2005). ESOM-Maps: tools for clustering, visualization, and classification with Emergent SOM. *Technical Report Dept. of Mathematics and Computer Science*. University of Marburg, Germany, No. 46.

- Ultsch A. & Siemon H. (1990). Kohonen's self-organizing neural networks for exploratory data analysis. *Proceedings of the International Neural Network Conference*. Kluwer, Paris.
- Vesanto J. (1999). SOM-based data visualization methods. *Intelligent Data Analysis*, 111 - 126.
- Vesanto J., Himberg J., Alhoniemi E. & Parhankangas, J. (20 de Abril de 2000). *SOM Toolbox for Matlab 5*. Technical Report A57, Helsinki University of Technology, Neural Networks Research Centre.

## 7. ANEXOS

### 7.1. CÁLCULO DO ÍNDICE DE HURST

A estatística R/S foi proposta por Hurst (1951) ao estudar o problema de dimensionamento de represas, onde procurou determinar a capacidade do reservatório ideal. Para tanto analisou os registos de décadas da vazão anual do rio Nilo.

O objetivo era determinar os volumes máximos e mínimos do reservatório, pois se pretendia evitar-lhe o transbordamento ou a secagem. A diferença entre os volumes máximos e mínimos seria o *range* de volumes do reservatório. Ainda era calculado o desvio padrão das vazões de água, sendo a estatística R/S o resultado da divisão do *range* pelo desvio-padrão, que é um valor adimensional. Ao analisar esta estatística para diversos períodos de tempo da série de registos, Hurst identificou que havia uma função que relacionava o valor da estatística R/S ao número de observações que entraram no cálculo. O resultado desta relação era um expoente “H” que mais tarde ficou conhecido como expoente ou índice de Hurst. Mais tarde Hurst verificou que a relação da estatística R/S também válida para outros fenómenos naturais.

Mandelbrot & Van Ness (1968) verificaram que o índice de Hurst aplicava-se também às séries que descrevem o movimento browniano fracionário, isto significa que a estatística R/S e o índice de Hurst podem ser utilizados para representar a presença de memória em séries com movimento browniano fracionário.

Os valores do índice de Hurst têm a seguinte interpretação:

- $0 < H < 0,5$ : série anti-persistente
- $H = 0,5$ : série aleatória (*Random Walk*)
- $0,5 < H < 1$ : série persistente

A seguir apresenta-se os cálculos para determinação do índice de Hurst.

1. Seja  $X_t$  o preço mínimo do índice DJI num determinado dia e  $r_t$  o retorno logarítmico deste preço, dado por:

$$r_t = \ln((X_t) / (X_{t-1}))$$

A série de  $N$  observações é dividida em  $\tau$  blocos de  $n$  elementos, onde  $1 \leq i \leq \tau$ . Em cada bloco  $i$  os elementos  $r_{t,i}$  são numerados com  $1 \leq t \leq n$ .

2. Calcula-se a média dos retornos de cada bloco

$$\bar{r}_i = \frac{1}{n} \sum_{t=1}^n r_{t,i}$$

3. Calcula-se o desvio-padrão em cada bloco:

$$S_i = \left[ \frac{1}{n} \sum_{t=1}^n (r_{t,i} - \bar{r}_i)^2 \right]^{1/2}$$

4. Calcula-se para cada bloco  $i$  a estatística  $(R/S)_i$ :

$$(R/S)_i = \frac{1}{S_i} \left[ \max_{1 \leq t \leq n} \sum_{K=1}^t (r_{k,i} - \bar{r}_i) - \min_{1 \leq t \leq n} \sum_{K=1}^t (r_{k,i} - \bar{r}_i) \right]$$

5. Calcula-se a média dos valores  $(R/S)_i$  e associa-se ao tamanho do bloco  $n$ :

$$(R/S)_n = \frac{1}{\tau} \sum_{i=1}^{\tau} (R/S)_i$$

6. Após os cálculos da estatística  $R/S$  para todos os blocos, obtém-se o índice de Hurst “ $H$ ” a partir da relação:

$$(R/S)_n = Cn^H$$

7. Dados os pares  $(R/S)_n$  e  $n$  obtidos, uma das formas de se estimar o índice de Hurst é aplicar a regressão OLS obtendo o logaritmo da série:

$$\log_{10}(R/S)_n = \log_{10}(C) + H \log_{10}(n)$$

Onde  $C$  é uma constante. Neste trabalho foi utilizada a contante  $\frac{\pi}{2}$ .

## 7.2. PACKAGE ORACLE UTILIZADO PARA CÁLCULO DAS VARIÁVEIS

As variáveis com cálculos mais complexos foram calculadas com auxílio de procedimentos PL/SQL, descritos abaixo. As demais variáveis foram calculadas de acordo com as fórmulas indicadas no capítulo 3. Metodologia.

```

CREATE OR REPLACE PACKAGE "PACK_STATISTICS" AS

    -- Declaração de Constantes
    pi CONSTANT NUMBER := 3.14159265358979;

    -- Declaração de Types
    TYPE typ_tab_hurst IS TABLE OF hurst_temp%ROWTYPE INDEX BY BINARY_INTEGER;

    TYPE typ_tab_sequencia IS TABLE OF sequencia_temp%ROWTYPE INDEX BY
    BINARY_INTEGER;

    -- Procedure para cálculo do índice Hurst
    PROCEDURE pro_calc_hurst(p_tab_hurst      IN OUT pack_statistics.typ_tab_hurst,
                           p_nro_elementos IN INTEGER);

    -- Procedure para calcular a variação diária
    PROCEDURE calc_capital_flutuation(p_id_stock VARCHAR2);

    -- Procedure para calcular o volume medio (volume(t)/volume medio(ndias)) de
    acordo com o nro de dias informado
    PROCEDURE calc_media_volume(p_sequencia IN OUT
    pack_statistics.typ_tab_sequencia,
                           p_ndias      NUMBER,
                           p_id_stock   VARCHAR2);

END pack_statistics;
/

CREATE OR REPLACE PACKAGE BODY "PACK_STATISTICS" AS

    /** Procedure para cálculo do índice Hurst
    * @param p_tab_hurst      Tabela auxiliar com o ID e PREÇO preenchidos
    * @param p_nro_elementos  Intervalo em dias para cálculo do hurst
    */
    PROCEDURE pro_calc_hurst(p_tab_hurst      IN OUT pack_statistics.typ_tab_hurst,
                           p_nro_elementos IN INTEGER) AS
        -- Cursor da média de um bloco
        CURSOR c_media_bloco(p_seq_ini NUMBER, p_seq_fim NUMBER) IS
            SELECT AVG(h.value) media
            FROM hurst_temp h
            WHERE h.id BETWEEN p_seq_ini AND p_seq_fim;

        -- Cursor com o desvio padrão e máximo e mínimo da variância
        CURSOR c_dev_max_min(p_seq_ini NUMBER, p_seq_fim NUMBER) IS
            SELECT STDDEV(h.value) desvio_padrao,
                   MAX(h.sum) maximo,
                   MIN(h.sum) minimo
            FROM hurst_temp h
            WHERE h.id BETWEEN p_seq_ini AND p_seq_fim;

        -- Variáveis
        v_qtde_blocos   INTEGER;
        v_seq_ini        INTEGER DEFAULT 0;
        v_seq_fim        INTEGER DEFAULT 0;

```



```

v_media_bloco    NUMBER(16, 4);
v_desvio_padrao  NUMBER(16, 4);
v_maximo         NUMBER(16, 4);
v_minimo         NUMBER(16, 4);
v_rs            NUMBER(16, 4);
v_hurst         NUMBER(16, 4);
BEGIN

    -- Insere registos na tabela temporária
    FORALL i IN p_tab_hurst.first .. p_tab_hurst.last
        INSERT INTO hurst_temp VALUES p_tab_hurst (i);

    FOR i IN p_tab_hurst.first .. p_tab_hurst.last LOOP

        v_seq_ini := (p_tab_hurst(i).id - p_nro_elementos);
        v_seq_fim := p_tab_hurst(i).id;

        OPEN c_media_bloco(v_seq_ini, v_seq_fim);
        FETCH c_media_bloco
            INTO v_media_bloco;
        CLOSE c_media_bloco;

        -- Preenche a coluna auxiliar com a variação de cada elemento em relação a
        média do bloco
        FOR n IN v_seq_ini .. v_seq_fim LOOP
            UPDATE hurst_temp
                SET VARIANCE = VALUE - v_media_bloco,
                    SUM      =
                        (VALUE - v_media_bloco) + CASE
                            WHEN n = v_seq_ini THEN
                                0
                            ELSE
                                (SELECT SUM FROM hurst_temp WHERE id = n - 1)
                        END
                WHERE id = n;
        END LOOP;

        -- Calcula-se o desvio padrão do bloco e o máximo e mínimo da variância
        OPEN c_dev_max_min(v_seq_ini, v_seq_fim);
        FETCH c_dev_max_min
            INTO v_desvio_padrao, v_maximo, v_minimo;
        CLOSE c_dev_max_min;

        -- Calcula-se a estatística R/S
        v_rs := (v_maximo - v_minimo) / v_desvio_padrao;

        dbms_output.put_line('Estatística R/S bloco' || i || ': ' || v_rs);

        v_hurst := log(10, v_rs)/(log(10, (pi / 2)) + log(10, p_nro_elementos));

        dbms_output.put_line('Hurst ' || i || ': ' || v_hurst);

        p_tab_hurst(i).result := v_hurst;

    END LOOP;
END pro_calc_hurst;

```

```

/** Procedure para calcular o volume medio (volume(t)/volume medio(ndias))
** de acordo com o nro de dias informado
* @param p_sequencia Tabela auxiliar com o ID e VOLUME preenchidos
* @param p_ndias Intervalo em dias para cálculo da média
* @param p_id_stock ID do índice DJI
*/
PROCEDURE calc_media_volume(p_sequencia IN OUT pack_statistics.typ_tab_sequencia,
                           p_ndias      NUMBER,
                           p_id_stock    VARCHAR2) IS

    v_media NUMBER;
BEGIN
    FOR i IN p_sequencia.first .. p_sequencia.last LOOP
        -- Media dos ultimos p_ndias
        BEGIN
            SELECT AVG(sm.px_volume)
            INTO v_media
            FROM stock_market sm
            WHERE sm.id_stock = p_id_stock
            AND sm.seq_stock BETWEEN (p_sequencia(i).id - p_ndias) AND
p_sequencia(i).id;
        EXCEPTION
            WHEN OTHERS THEN
                v_media := p_sequencia(i).value;
        END;
        --
        p_sequencia(i).result := p_sequencia(i).value / v_media;
    END LOOP;
EXCEPTION
    WHEN OTHERS THEN
        NULL;
    END calc_media_volume;

/** Procedure para calcular a variação do preço diário.
** px low(t)/px low(t-1)
* @param p_id_stock ID do índice
*/
PROCEDURE calc_capital_flutuation(p_id_stock VARCHAR2) IS
    CURSOR c_stock IS
        SELECT *
        FROM stock_market
        WHERE id_stock = p_id_stock
        ORDER BY seq_stock;

    v_capital_flutuation NUMBER;
BEGIN
    FOR r IN c_stock LOOP
        BEGIN
            SELECT r.px_low / sm.px_low
            INTO v_capital_flutuation
            FROM stock_market sm
            WHERE sm.id_stock = p_id_stock
            AND sm.seq_stock = (r.seq_stock - 1);
        EXCEPTION
            WHEN OTHERS THEN

```

```

        v_capital_flutuation := 1;
    END;

    UPDATE stock_market m
        SET m.ix_price_fluctuation = v_capital_flutuation
        WHERE m.id_stock = r.id_stock
        AND m.seq_stock = r.seq_stock;
    END LOOP;
EXCEPTION
    WHEN OTHERS THEN
        dbms_output.put_line(SQLERRM);
    END calc_capital_flutuation;

END pack_statistics;

```